PROJETO DE GRADUAÇÃO

O *BIG DATA* COMPORTAMENTAL COMO FERRAMENTA DE PESQUISA NA ENGENHARIA DE PRODUÇÃO: ALINHAMENTO METODOLÓGICO E FERRAMENTAL

Por

Ana Bárbara Pereira Plá

Brasília, 09 de julho de 2019

UNIVERSIDADE DE BRASÍLIA

FACULDADE DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO

UNIVERSIDADE DE BRASÍLIA
Faculdade de Tecnologia
Departamento de Engenharia de Produção

PROJETO DE GRADUAÇÃO

O *BIG DATA* COMPORTAMENTAL COMO FERRAMENTA DE PESQUISA NA ENGENHARIA DE PRODUÇÃO: ALINHAMENTO METODOLÓGICO E FERRAMENTAL

Por

Ana Bárbara Pereira Plá

Relatório submetido como requisito parcial para obtenção do
grau de Engenheira de Produção

Banca Examinadora

Prof. Ari Melo Mariano, Ph.D. – UnB/EPR (Orientador) _____

Prof. Msc. Roberto Ávila Paldês _____

Prof. Msc. Maíra Rocha Santos _____

Brasília, 09 de julho de 2019

RESUMO

Este trabalho tem como objetivo apresentar uma construção metodológica para o uso do *Big Data* Comportamental (BDC) na pesquisa científica em Engenharia de Produção. No *Big Data* Comportamental os dados são influenciados pelos sujeitos que estão sendo analisados, que interagem continuamente e de forma consciente com os dados e tem um papel ativo na pesquisa. O BDC permite que se entenda as ações e o comportamento humano em um nível que não era possível antes, com a possibilidade de se transformar as relações das organizações e da sociedade. A pesquisa é exploratória com abordagem qualitativa e quantitativa, com a revisão da literatura por meio da metodologia da Teoria do Enfoque Meta Analítico Consolidado (TEMAC). Foram encontrados 6 resultados na base de dados *Web of Science* para o termo. Identificou-se a necessidade de uma metodologia e uma ferramenta que garantam o rigor científico necessário e que se adequem às necessidades do BDC. Foi escolhida a metodologia *Design Science Research* (DSR) e a ferramenta *Partial Least Square Structural Equation Modeling* (PLS-SEM). Essa escolha se deu pela abordagem de artefatos de ambas, que se aproxima do BDC. Além disso, elas permitem a realização de análises prescritivas, que são uma necessidade cada vez maior das organizações para fundamentar decisões.

Palavras-chave: Big Data Comportamental, Design Science Research, Análise Prescritiva, PLS-SEM.

ABSTRACT

This paper aims to present a methodological framing for the use of Behavioral Big Data (BBD) in scientific research in Production Engineering. In the Behavioral Big Data data are influenced by the subjects being analyzed, which interacts continuously and in a conscious way with the data and has an active role in the research. The BBD allows actions and behaviors to be understood in a level that was not possible before, with the possibility to change the relationship between organizations and society. It has been an exploratory research with a qualitative and quantitative approach, with a review of the literature by the TEMAC method. Six articles were found in the database Web of Science for the term. It was identified the need for a methodology and a tool that ensure the scientific rigor and adapt to the needs of the BBD. It was chosen the Design Science Research methodology and the Partial Least Square Structural Equation Modeling tool (PLS-SEM). This choice was convenient because of the artifact approach of them that gets close with the BBD. Besides, they allow prescriptive analysis, an increasing demand of the organizations to justify their decisions.

Keywords: Behavioral Big Data, Design Science Research, Prescriptive Analytics, PLS-SEM.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

A quarta revolução industrial, assim como as suas antecessoras, surge causando um grande impacto e transformação para a sociedade, trazendo consigo uma grande oportunidade de crescimento para os países. Enquanto a primeira revolução industrial trouxe uma grande ruptura na estrutura mecânica dos meios de produção, a segunda no âmbito elétrico, com a chegada da produção em massa, e a terceira focou nas mudanças trazidas pelos processos de automação, a quarta revolução industrial, através da utilização de inteligência artificial, robótica, *Big Data*, entre outras tecnologias, busca uma junção entre as tecnologias físicas, digitais e biológicas. Segundo Schwab (2019), a crescente harmonização e integração de descobertas nas diferentes áreas é o que torna a quarta revolução industrial única, e inovações que resultam desta colaboração de tecnologias de áreas diversas já são reais e estão transformando a sociedade.

Os avanços dessa nova revolução chegam em uma velocidade sem precedentes e impactam as indústrias mundialmente. Nas fábricas os processos são integrados com tecnologias digitais, tornando-os cada vez mais independentes. Segundo Motta et al. (2018) a digitalização, é a palavra que define essa nova revolução, e a aderência dela à indústria "resultou na manufatura avançada, e como consequência, do maquinário interconectado em rede e junção do mundo físico e virtual, que é caracterizada pela integração e pelo controle dos processos de produção em toda a cadeia logística".

A Indústria 4.0 surge então como a indústria que engloba essas inovações tecnológicas aplicadas na manufatura, tornando os processos automatizados e eficientes. Uma das premissas da Agenda da Indústria 4.0 é "propor agenda centrada no industrial/empresário, conectando instrumentos de apoio existentes, permitindo uma maior racionalização e uso efetivo, facilitando o acesso dos demandantes, levando o maior volume possível de recursos para a "ponta"". (MINISTÉRIO DA INDÚSTRIA, COMÉRCIO E SERVIÇOS, 2019) Ou seja, a utilização de dados é uma das bases para que esta evolução seja possível.

Além disso, a Indústria 4.0 se baseia em 9 pilares centrais: Sistemas Ciberfísicos, Internet das Coisas, Manufatura Aditiva, Manufatura Digital ou Simulação, Fábricas Inteligentes, Análise de Grandes Quantidades de Dados, Computação em Nuvem, Segurança Digital e Robótica Avançada. Um fator que permeia os pilares da indústria 4.0 são as grandes

quantidades de dados. Em meio a tanto desenvolvimento tecnológico há um grande volume de produção de dados e informações.

Além da grande quantidade, esses dados apresentam uma grande variedade. Dentre todos os desafios que chegam com a Indústria 4.0, as empresas e indústrias enfrentam a imprevisibilidade do mercado e dos clientes, sendo cada vez mais necessário antecipar tendências para basear as decisões tomadas.

Existe, portanto, uma demanda pelo tratamento e análise desses dados, especialmente com o objetivo de auxiliar na tomada de decisão. O *Big Data* surge, então, como a interpretação desta grande quantidade de dados, entregando conhecimento para as pessoas interessadas. Lohr (2012) descreve o termo como avanços tecnológicos que possibilitam novas abordagens no entendimento do mundo e no processo de tomada de decisões.

Apesar do desenvolvimento das tecnologias e do avanço das ferramentas de análise de dados contribuírem para a integração dos dados e tratamento para tomada de decisões, Shmueli (2017b) explica que novos desafios aparecem decorrentes da limitação dos dados tradicionais, inanimados (vindos da engenharia) e fisiológicos (vindos da medicina), e com o surgimento dos dados comportamentais, em que os sujeitos que estão sendo analisados tem uma interação contínua e consciente, enriquecendo os dados com emoções, intenções, decepções entre outros aspectos humanos e sociais. Nesse contexto surge a ideia de *Big Data* Comportamental (BDC) (*Behavioral Big Data*).

Apesar de todo o potencial que o BDC apresenta, seu uso ainda não é muito disseminado e não são encontrados muitos estudos sobre o tema. Além disso, as vantagens do uso do BDC na análise de dados e aplicações na tomada de decisão não são encontradas facilmente e esclarecidas entre as pessoas, se tornando desta forma, algo que não é utilizado com frequência. E quando se trata de estudos científicos a situação é ainda mais complexa, com escassos resultados em bases de dados importantes como *scopus*, que apresenta apenas 7 artigos e *Web of Science*, com 6 registros.

Sabe-se que existe, cada dia mais, uma quantidade maior de dados sobre os mais variados assuntos que se possa desejar, porém, ainda é desconhecido o diferencial de se utilizar o *Big Data* Comportamental para auxiliar os centros de pesquisa e organizações. Além disso, o surgimento de novas técnicas de análise de dados e a possibilidade de sua adesão na pesquisa científica revela uma necessidade de alinhamento com uma metodologia e sua execução.

1.1 Problema da Pesquisa

Tendo em vista as informações apresentadas, este trabalho tem como foco responder à seguinte questão: Como o *Big Data* Comportamental pode se inserir no contexto da pesquisa científica em Engenharia de Produção?

1.2 Justificativa

Esse trabalho se justifica no âmbito social pela relevância que o assunto *Big Data* Comportamental tem nos mais variados contextos da sociedade, além de ter um grande potencial de apresentar novos resultados a partir dos dados disponíveis e melhorar a competitividade de empresas e de impactar nas decisões delas, de governos e de pessoas.

No âmbito científico se justifica pela possibilidade de incrementar a pesquisa com ferramentas mais robustas, auxiliando na obtenção de resultados mais efetivos e contribuindo com a discussão do assunto, ainda em crescimento.

A área de Engenharia de Produção, em uma de suas definições mais fundamentais, busca solucionar problemas e otimizar processos, seja em sua aplicação diária, seja na pesquisa científica. Assim, contribuir no desenvolvimento deste tema colabora na melhoria do campo, pelo potencial que oferece para suporte de decisões e resolução de problemas, sendo essencial a utilização de dados desse tipo nas mais diferentes atuações de um engenheiro de produção.

1.3 Objetivos

O objetivo geral deste trabalho é apresentar uma construção metodológica para o uso do *Big Data* Comportamental na pesquisa científica em Engenharia de Produção.

Para que se consiga atingir o objetivo geral proposto, são definidos os seguintes objetivos específicos:

- Delimitar o conceito de *Big Data* Comportamental;

- Pesquisar possíveis caminhos metodológicos para o uso do *Big Data* Comportamental na pesquisa científica;
- Mapear ferramentas que incorporem os conceitos de *Big Data* Comportamental.

1.4 Estrutura dos Capítulos

Este trabalho está estruturado da seguinte forma: O capítulo 2 aborda a revisão da literatura, realizada com base no enfoque meta-analítico. O capítulo 3 apresenta a revisão bibliográfica realizada acerca dos principais conceitos para a realização do trabalho. O capítulo 4 fala da metodologia usada no trabalho. No capítulo 5 é realizada uma análise acerca da metodologia e ferramenta sugeridas para o uso do *Big Data* Comportamental no contexto da Engenharia de P. O capítulo 6 apresenta as principais conclusões do trabalho.

2. REVISÃO DA LITERATURA

Apesar de não ser obrigatório, cada vez mais é esperado que se realize a revisão bibliográfica para a garantia de se encontrar materiais confiáveis para o trabalho, além de se assegurar que os principais autores de determinado assunto tenham sido estudados. O enfoque utilizado para a realização deste trabalho foi o modelo TEMAC – Teoria do Enfoque Meta Analítico Consolidado. Segundo Mariano (2017), ele garante o respeito para uma avaliação de qualidade do artigo. O TEMAC é composto por três fases: preparação da pesquisa com o uso de múltiplas bases de dados, apresentação e inter-relação dos dados e detalhamento e validação através de evidências. Este modelo pode ser observado na figura a seguir.

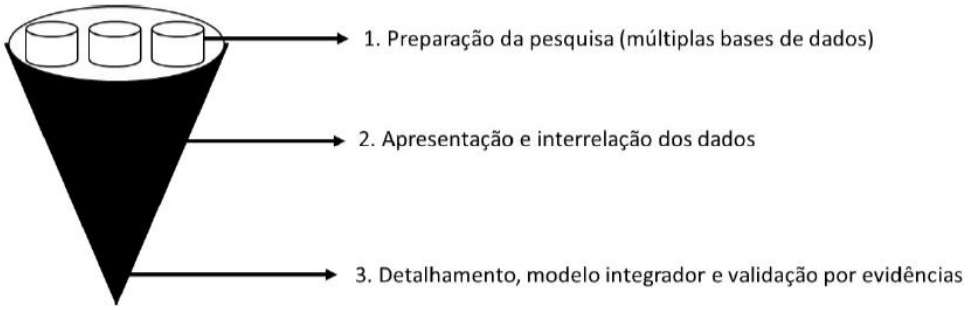

Figura 1 - TEMAC

Fonte: MARIANO et al., 2017.

2.1 Primeira Fase: Preparação da Pesquisa

Para a primeira fase do TEMAC foram utilizadas as bases de dados *Web of Science, Scopus* e *Google Scholar*. A primeira e a segunda foram escolhidas pelo seu reconhecimento e qualidade de informações e a terceira foi escolhida pela grande multidisciplinaridade de assuntos e maiores opções de resultados. O termo escolhido para a pesquisa foi "*Behavioral Big Data*".

Na *Web of Science*, sem limitação de tempo, foram encontrados 6 resultados, na *Scopus* foram encontrados 7 registros, e na *Google Scholar* foram encontrados 86 resultados para o termo. A pouca quantidade de trabalhos encontrados com esse termo indica que o assunto é pouco estudado, tendo ainda muito o que ser explorado sobre *Big Data* Comportamental.

Ainda na *Web of Science,* onde se encontram os artigos científicos sobre o tema, a principal autora encontrada foi Galit Shmueli com 2 publicações, ambas feitas no ano de 2017, o primeiro ano encontrado com o termo na base de dados. O principal artigo encontrado é o *Research Dilemmas with Behavioral Big Data*, que será melhor explicado nos próximos tópicos deste trabalho.

2.2 Segunda Fase: Apresentação e Interrelação dos Dados

A segunda etapa do TEMAC deve apresentar as relações entre as principais publicações encontradas nas bases de dados.

Nos textos encontrados na *Web of Science,* 2 deles são artigos e 4 são materiais editoriais. A autora mais citada é Galit Shmueli, com 7 citações no total, 4 do artigo *Analyzing Behavioral Big Data: Methodological, practical, ethical, and moral issues*, e 3 do artigo *Research Dilemmas with Behavioral Big Data*. Os outros artigos encontrados não possuem citações.

Ainda é possível verificar as universidades que realizam a maior quantidade de publicações. São elas: *National Tsing Hua University,* com duas publicações e, *City University of Hong Kong, Miami University, Reancon, Texas A M University System, Union College* e *University of San Francisco*, todas com uma publicação cada. Quanto à localidade dos

trabalhos, 3 deles foram feitos nos Estados Unidos, 2 em Taiwan e 1 na China. Os seis artigos encontrados foram publicados em inglês.

Para identificar as linhas de pesquisa mais exploradas relacionadas ao trabalho, foram coletadas todas as palavras-chave dos artigos utilizados para a realização do trabalho. A partir delas foi criada uma nuvem de palavras, através da ferramenta de análise *TagCrowd*, que permite que os dados encontrados sejam dispostos de uma forma mais visual. A Figura 2 abaixo mostra, então, as palavras que ocorreram com maior frequência, sendo evidenciado ainda pelo tamanho da fonte, que é proporcional à quantidade de citações de cada termo.

Figura 2 – Nuvem de Palavras

Fonte: Própria. Extraída do site *TagCrowd*.

A partir da Figura 2 é possível verificar que dentro do assunto abordado, as principais temáticas encontradas são: *data, analytics, behavior, science, business e intelligence*.

De uma maneira geral, os textos utilizados que contém as temáticas de *data, analytics, business* e *intelligence* tratam de aspectos mais técnicos e características gerais sobre dados e *Big Data*. *Behavior* e *science* evidenciam as temáticas acerca dos dados comportamentais e ciências, que analisam seus usos, metodologias e desafios em diversos âmbitos.

Além destes textos que possuem a finalidade de compreensão dos conceitos e problemas acerca do tema de análise de dados, ressalta-se também o aparecimento de termos menos óbvios no âmbito estudado, como PLS, inovação, regressões, estudos estatísticos e prescritivo. Essas palavras representam novas abordagens no contexto de dados, como estudos estatísticos que realizam regressões, como o PLS ou que possuem o objetivo de realizar análises prescritivas, que ainda são pouco observadas neste contexto.

2.3 Terceira Fase: Detalhamento, Modelo Integrador e Validação por Evidências

Nesta etapa busca-se análises mais profundas do que as anteriores, para que se possa selecionar os textos essenciais para o trabalho.

Primeiro, foi feita uma análise de *coupling* e *co-citation*, que identificam as relações entre autores e referências na literatura. A primeira análise, *coupling*, busca os artigos que possuem citações iguais. De forma semelhante, a segunda análise, de *co-citation*, verifica artigos que são citados juntos. Desta forma, "o *Coupling* traz uma perspectiva de frentes de pesquisa e a Co-citação das abordagens mais utilizadas". (MARIANO et al., 2017)

A análise de co-citações permite verificar quais artigos são citados em conjunto, o que sugere que possuem uma linha de pesquisa semelhante.

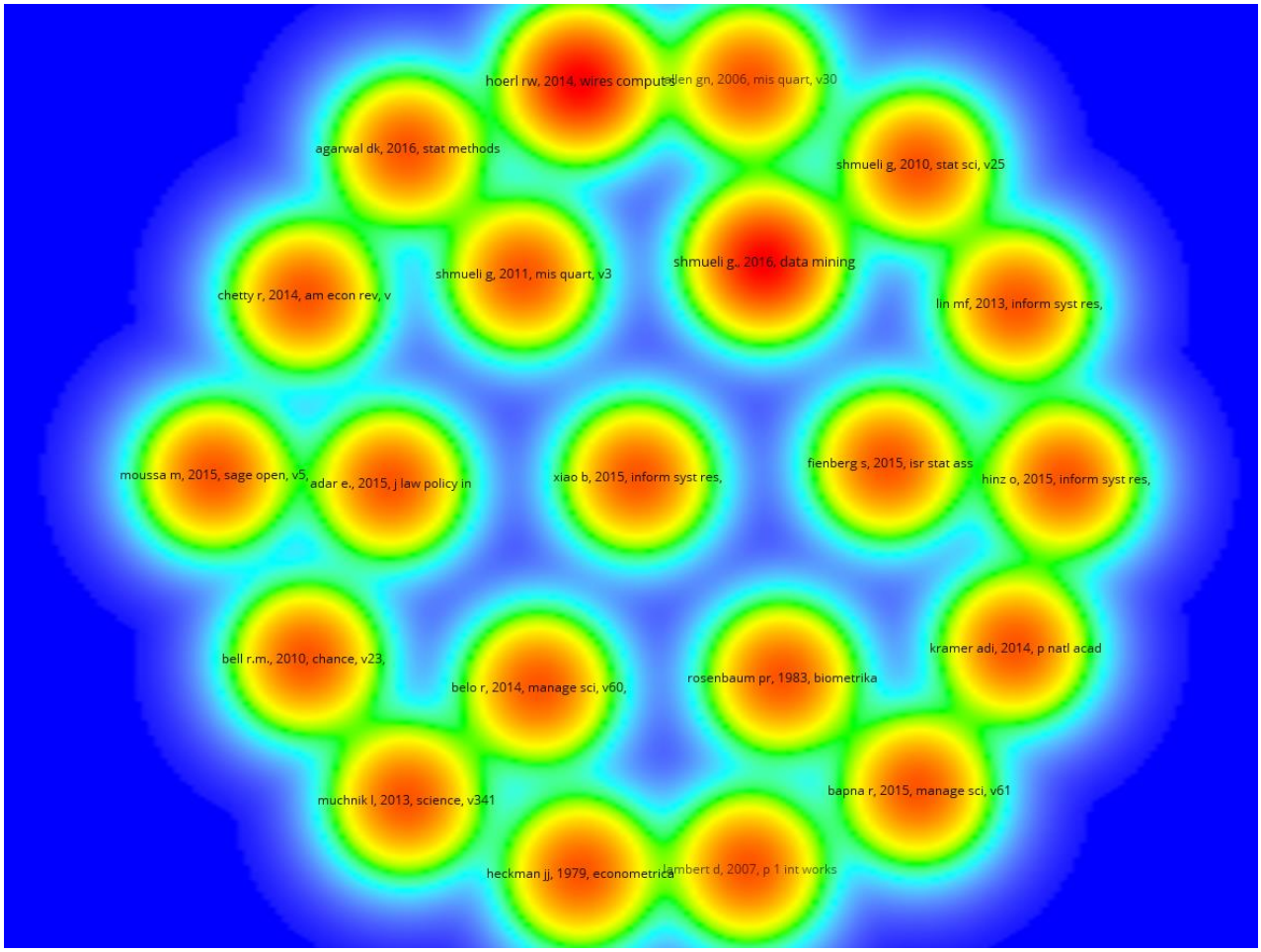

Figura 3 – Mapa de Calor de Co-citação

Fonte: própria. Extraído do programa *VOSviewer*.

A partir do mapa de calor representado na Figura 3, percebe-se um grande espaçamento entre os autores citados, além da não formação de ilhas que se destaquem das outras. Isso indica que não há no momento trabalhos semelhantes a partir desta perspectiva. Existem muitas abordagens diferentes que ainda estão isoladas quanto ao *Big Data* Comportamental.

Algumas delas, como a de Hoerl et al. (2014) visam mostrar aplicações de estudos com o uso de dados comportamentais. Neste caso, ele usa o pensamento estatístico para problemas de *Big Data*, o que, apesar de ainda não utilizar o nome *Big Data* Comportamental, se trata de exemplificação desse tipo de dado. Agarwal et al. (2016) também demonstram a aplicação de métodos estatísticos no contexto de dados atual, para sistemas de recomendação. Já Lin et al. (2013) buscam em sua pesquisa mostrar problemas relacionados ao uso de *p-value* no estudo de grandes quantidades de dados e como evitar esses problemas.

Outros deles, como o estudo de Xiao et al. (2015), buscam identificar desafios que se encontram ao trabalhar com dados comportamentais. Neste caso, são mostradas formas de detectar abordagens enviesadas em recomendações de produtos *online*. Já o trabalho de Bapna et al. (2015) mostra um experimento de influência de amigos em redes sociais, e as consequências que isso gera nos comportamentos dos indivíduos e, consequentemente, nos dados comportamentais destes. Muchnik et al. (2013) também buscam mostrar a influência que o comportamento de outras pessoas tem nas ações de indivíduos, neste caso, ações de avaliação de comentários. Kramer et al. (2014) do mesmo modo estudam como se dá a propagação de emoções no comportamento individual através do uso das redes sociais.

Já a análise de *Coupling*, que se refere ao acoplamento bibliográfico, apresenta os artigos que possuem referências em comum, mostrando as abordagens que estão se destacando atualmente.

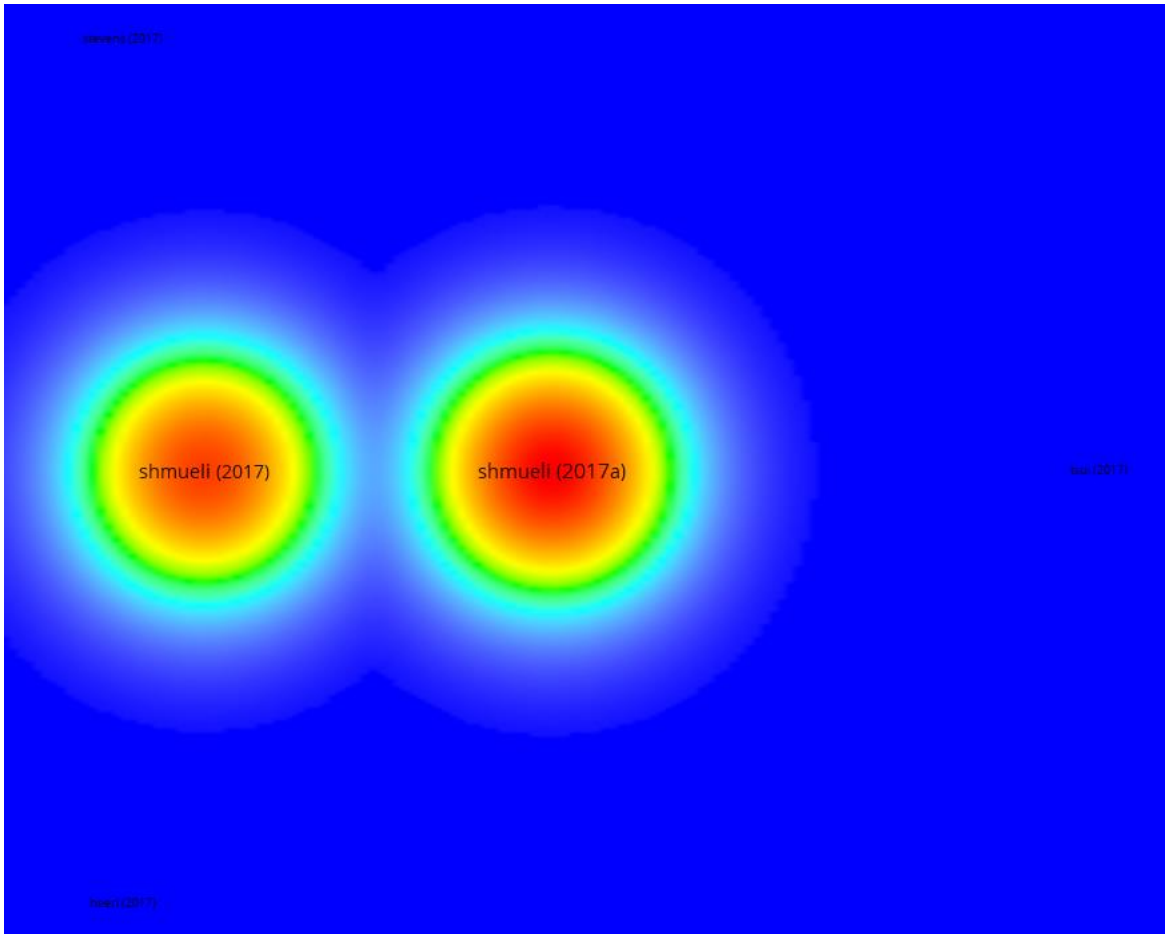

Figura 4 – Mapa de Calor de *Coupling*

Fonte: própria. Extraído do programa *VOSviewer*.

A partir do mapa de calor da Figura 4, percebe-se que o assunto no momento e no contexto de publicações científicas é dominado pela abordagem de Shmueli. Essa abordagem foca em definir o conceito de BDC, diferenciar esse tipo de dado de outros tipos de dados existentes e explicar seus principais desafios em diversos âmbitos, ressaltando que ainda faltam estudos para que se possa propor soluções para essas questões.

Finalmente, foi criada uma tabela com os principais autores e suas contribuições. Além de falar sobre os trabalhos encontrados na base de dados *Web of Science*, também se acrescentou os principais textos encontrados na *Google Scholar*, pois foram encontrados poucos textos na primeira base.

Tabela 1 – Principais artigos e suas contribuições

Título	Autores	Colaborações/Amostra
Research Dilemmas with Behavioral Big Data	Shmueli, G	A autora busca explicar o *Big Data* Comportamental, a relação entre pesquisador, dados, sujeitos e questão de pesquisa. Além disso, fala sobre dilemas enfretados no contexto de estudos de BDC para pesquisas de causas comportamentais.
Analyzing Behavioral Big Data: Methodological, practical, ethical, and moral issues	Shmueli, G	O trabalho descreve o *Big Data* Comportamental e avalia as oportunidades e dificuldades que surgem ao se aplicar abordagens estatísticas e de mineração de dados ao BDC.
Predictive Modeling with Big Data: Is Bigger Really Better?	Fortuny, EJ ; Martens, D; Provost, F	A partir da premissa de que o uso de dados tem sido usado para realizar análises preditivas, o trabalho demonstra que empresas que possuem maior quantidade de dados conseguem análises preditivas mais valiosas.
Data ex Machina: introduction to big data	Lazer, D; Radford, J	O trabalho conceitua *Big Data* e seus domínios. Além de defender o potencial que existe nesses dados para a observação de fenômenos, também evidencia vulnerabilidade e tendências existentes.
Trade offs, limitations and promises of Big Data in social science research	White, P; Breckenridge, RS;	Há um aumento nas pesquisas de *Big Data* no âmbito acadêmico, público e governamental. O artigo mostra as limitações e *trade-offs* nessas pesquisas, o que é essencial para um melhor entendimento do termo no longo prazo.
Applying statistical thinking to 'Big Data' problems	Hoerl, RW; Snee, RD	Com o aumento no uso do *Big Data*, algumas teorias sugerem que fundamentos estatísticos não são mais fundamentais para a análise de dados, mas o trabalho busca mostrar a relevância desses fundamentos no âmbito do *Big Data.*

Título	Autores	Colaborações/Amostra
Mining the quantified self: personal knowledge discovery as a challenge for data science	Fawcett, T	O autor apresenta oportunidades e desafios trazidos com o *quantified self*, que diz respeito a pessoas comuns gravando e analisando aspectos da sua vida com o objetivo de entender e melhorar a si mesmas.
Using behavioral big data for public purposes: Exploring frontier issues of an emerging policy arena	Saramajiva, R; Lokanathan, S;	O trabalho analisa questões de privacidade na pesquisa e aplicações do *Big Data* no âmbito público. O *Big Data* analisado é gerado de dados advindos de comportamentos humanos obtidos com o auxílo de tecnologias utilizadas diariamente.

Fonte: própria

Por fim, buscou-se por contribuições, metodologias e ferramentas que pudessem respaldar o uso do BDC na pesquisa científica. A partir disso, foi criado uma figura com o principal resultado.

Figura 5 – Metodologias para o BDC.

Fonte: Própria.

De uma maneira geral, os estudos que utilizam a pesquisa com BDC ainda são recentes, e objetivam comprovar hipóteses e suposições antigas, mas que ainda não tinham como ser provadas, principalmente pela falta de dados que pudessem testar essas ideias. Desta forma, os estudos encontrados que exemplificam usos de BDC são estudos exploratórios, pois eles testam hipóteses em problemas que ainda são pouco conhecidos, como é o caso de problemas com o uso de BDC.

Esse tipo de estudo requer uma metodologia específica e comum entre os estudos, o que não foi encontrado, apesar de existir certa similaridade entre os estudos de BDC atuais. O *Design Science Research* (DSR) surge, então, como uma opção para esse tipo específico de dado, com o apoio da ferramenta PLS-SEM, que compartilha dos seus princípios e características.

3. REFERENCIAL TEÓRICO

3.1 Dados

Um dado é um elemento bruto que, em conjunto, forma uma informação que produz conhecimento e serve de base para tomadas de decisões. Em um mundo em que as pessoas têm cada vez mais acesso à tecnologia e interação umas com as outras, os dados são gerados em uma proporção não observada anteriormente.

Além da grande quantidade de dados que vem sendo disponibilizada, é cada vez maior o saber que os computadores e tecnologias conseguem gerar com estas informações. Agarwal e Dhar (2014) atentam que os novos problemas e questionamentos mais desafiadores que surgem na sociedade induzem a criação de melhores algoritmos e sistemas para conseguirem lidar com todas essas novidades.

Com os avanços cada vez maiores no contexto de dados, as decisões passaram a ser, progressivamente, baseadas em análises de dados ao invés dos instintos e experiências de cada um. Shmueli (2017a) alerta que governos e empresas já perceberam que dados sobre comportamento humano podem ajudá-los a tomar decisões melhores.

Embora o uso dos dados esteja em alta no processo de decisão, seu uso sistemático é registrado na literatura científica desde 1970 (Figura 6).

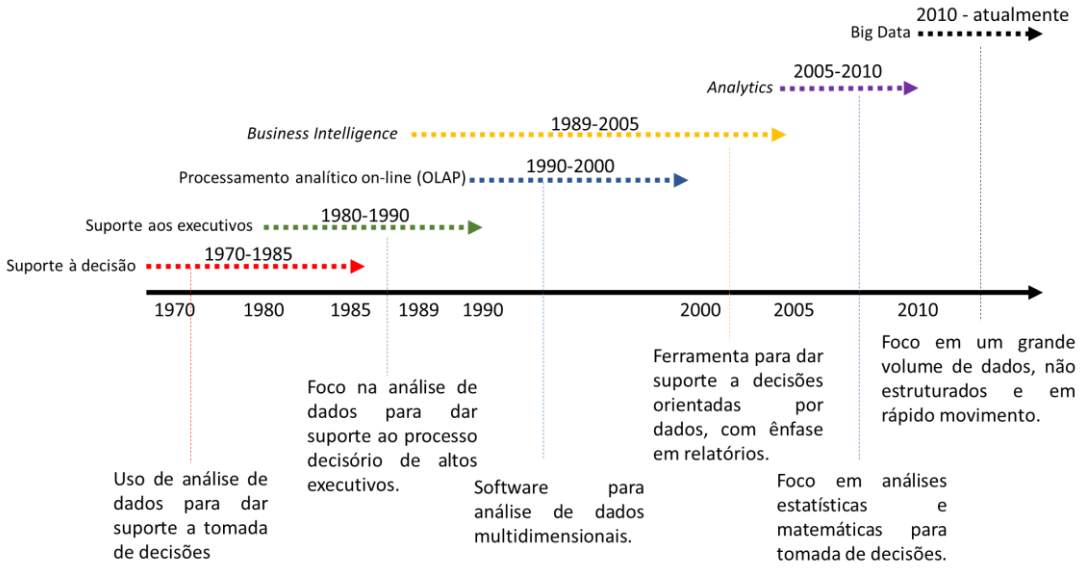

Figura 6 – Gerações de Análises de Dados.
Fonte: Goldschmidt et al. (2015).

Mesmo tendo sua origem sistemática com o uso da tecnologia e computação desde 1970, até a chegada do *Business Intelligence* (BI), o processo de tratamento era específico, especializado e restrito para algumas áreas organizacionais. A partir do BI, a cultura do tratamento e uso dos dados passa a ser ampliada a toda organização.

Geralmente *Business Intelligence* é definido como "técnicas, tecnologias, sistemas, práticas, metodologias e aplicações que analisam dados críticos dos negócios para auxiliar organizações a entender melhor seu negócio e mercado para tomar decisões" (CHEN et al., 2012). Chaudhuri (2011) define *Business Intelligence* como um conjunto de tecnologias que oferece suporte para as empresas tomarem melhores decisões de forma mais rápida.

Segundo Negash (2004), BI auxilia nos processos de decisão puxando informações de outros sistemas, sendo uma atividade precisa, o que faz com que a informação seja relevante e entregue no tempo em que é necessário o uso desta.

As técnicas, ferramentas e métodos que integram o BI geralmente são baseadas em métodos estatísticos ou em técnicas de *data mining* desenvolvidas no século passado, que geram análises, segmentação de dados e *clustering*, análises de classificação e regressão, detectam anomalias e geram modelos preditivos (CHEN et al., 2012).

Embora grande parte das empresas brasileiras ainda estejam criando uma maturidade para usar os dados a nível de BI, o processo de tratamento e uso dos dados avançou por meio de um novo conceito, o *Data Analytics* (Figura 6).

Data Analytics significa análise de dados para gerar informações com um objetivo. Isso é feito a partir da busca de padrões com o uso de algoritmos, técnicas e ferramentas para realizar a análise dos dados disponíveis. Segundo Chen et al. (2012) *Analytics* são as formas de se adquirir dados, processá-los para obter padrões e *insights* e entregar resultados para os interessados. Para Hoerl et al. (2014) são métodos quantitativos que buscam informações significativas em dados.

O *Data Analytics* começa com a origem dos dados, que podem ser obtidos através de bancos de dados, coletas manuais, sites, etc. Dentro desta parte, existe o conceito de *data warehousing*, que se refere ao transporte de dados de diversas fontes para um sistema integrado, chamado de *data warehouse*, que pode ser traduzido como um armazém de dados.

Depois é necessário ser feito o tratamento destes dados, que muitas vezes não estão da forma desejada para análise. Após o tratamento dos dados é possível encontrar o procedimento

mais conhecido como *data mining*, a mineração de dados. A mineração de dados se refere à extração de dados, buscando padrões e semelhanças que possam ser usados para basear teorias e tomadas de decisões. O *data mining* é importante especialmente quando se lida com grandes quantidades de dados, pois muitas vezes ele é o responsável por encontrar padrões e relações que poderiam nunca ser percebidos dentro da imensidão dos dados (FAWCETT, 2015).

Então, quando os dados já foram tratados e o *data mining* executado, é realizada a análise destes dados, que é feita a partir das questões que geraram o estudo. Por fim pode ser realizado um processo de validação e criação de conhecimento para o melhor aproveitamento e cumprimento dos objetivos deste processo.

Finalmente o estágio atual da abordagem sobre os dados, o *Big Data*. *Big Data* é um conjunto de dados integrados e processos de análise para prover informações para melhores tomadas de decisões (AKTER et al., 2017). Em outras palavras, *Big Data* é um grande conjunto de dados, informações de clientes, publicações em redes sociais, entre outros, que possibilitam uma interpretação e análise que podem, a partir de uma reformulação de estratégias, criar novas oportunidades de negócios (KUAZAQUI, 2018).

White e Breckenridge (2014) definem *Big Data* como um conjunto de dados grande, diverso e complexo, gerado a partir de diferentes intrumentos, sensores e operações realizadas na internet e outras fontes digitais disponíveis atualmente ou no futuro, que podem gerar progresso nas ciências e inovações, incentivar a criação de ferramentas e *softwares* e aumentar a compreensão de processos e interações humanas e sociais, além de promover melhorias para a sociedade. Segundo Cook e Forzani (2018), o *Big Data* promete revelar conexões entre os dados ou variáveis que podem levar a novos conhecimentos científicos e melhoria de processos.

Embora o *Big Data* esteja relacionado a “grandes dados”, etimologicamente, é comum usar o termo como sinônimo de *Big Data Analytics*, que, em verdade é o uso de algoritmos de análise em plataformas para encontrar padrões desconhecidos ou escondidos no *Big Data*. Segundo Gandomi e Haider (2015) *Big Data Analytics* são técnicas e ferramentas usadas para realizar análise e extrair inteligência do *Big Data.*

Segundo Hu (2014) o *Big Data Analytics* pode ser classificado em dois paradigmas distintos: Processamento *Streaming*, que assume que o valor associado ao dado depende de quão novo o dado é, e Processamento *Batch*, que primeiro armazena os dados para então analisá-los.

Big Data e *Data Analytics* se complementam. *Big Data* sem *Analytics* é apenas uma grande quantidade de dados desorganizados, e *Analytics* sem *Big Data* são apenas ferramentas e técnicas sem nenhuma aplicação (SANDERS, 2016). Transformar dados em previsões futuras é uma das propostas mais importantes do *Big Data Analytics*. Enquanto na estatística o principal é o modelo, no *Big Data Analytics* o principal são os dados (ARADAU e BLANKE, 2017).

Deste modo, referir-se a *Big Data* como *Big Data Analytics* é incorreto, porém seu uso regular tornou o termo aceito.

Não é exagero dizer que o *Big Data* é uma das principais revoluções tecnológicas nos sistemas econômicos e acadêmicos desde o surgimento da internet e da economia digital. Os avanços obtidos com as novas tecnologias permitem uma captura cada vez mais rápida dos dados, além de existir uma disponibilidade enorme e complexa destes dados, que geram oportunidades de estudos sobre os mais variados assuntos. O *Big Data* "ainda pretende entregar a informação certa para a pessoa certa no tempo certo e da forma certa, mas agora consegue fazer isso de forma significantemente mais sofisticada" (AGARWAL e DHAR, 2014).

Segundo Martin (2015) o *Big Data* se diferencia da análise tradicional de dados pela variedade de dados combinados, que no *Big Data* vem de uma grande variedade de fontes usadas de forma inovadora. O *Big Data* possui dados que tem muita variabilidade e complexidade e utiliza técnicas e tem aplicações tão complicadas e variadas que demandam novas tecnologias para suportar toda essa quantidade de informações. Gandomi e Haider (2015) explicam que, nesse contexto, a variabilidade se refere à variação no fluxo dos dados, enquanto a complexidade se deve ao fato de os dados serem gerados em fontes diversas.

Além do tamanho, o *Big Data* também pode apresentar outras características marcantes como: ser relacional – relaciona informações entre indivíduos e grupo; agregada – combina dados e variáveis que provém de diversas fontes e sensores; multinível – combinando medidas em um nível individual e de grupo; e mista – que são originados de fontes diferentes (WHITE e BRECKENRIDGE, 2014). Boyd e Crawford (2011) reforçam essas características ao dizer que a capacidade do *Big Data* de se relacionar com outros dados é um dos seus aspectos mais notáveis, e que seu valor se dá através de padrões que são observados a partir das conexões entre dados e as relações obtidas entre indivíduos e grupos.

A promessa do *Big Data* é a detecção de informações ocultas e padrões desconhecidos que possibilitam a predição de eventos futuros, como ataques terroristas, crimes e crises migratórias, por exemplo (ARADAU e BLANKE, 2017).

Apesar de o *Big Data* ter um grande potencial para o estudo preditivo, Agarwal e Dhar (2014) destacam que existe uma discussão sobre se o *Big Data* pode ser usado apenas em casos de predições ou se também abrange o entendimento das relações causais nos resultados observados nas pesquisas, existindo então, um dilema entre predição e explicação no uso do *Big Data*, pois seu uso é tão possível para criar hipóteses quanto para testá-las e entender suas causas. Contudo, vale destacar que alguns domínios enxergam o *Big Data* como mais valioso no âmbito da predição, pois em algumas situações, por exemplo, é necessário se tomar decisões rápidas para evitar prejuízos, e nestas situações, a predição acaba sendo mais valiosa e rápida do que o estudo das possíveis causas, o que não significa que o estudo das relações causais não seja importante também.

Além disso, como diz Junqué de Fortuny et al. (2013), a mentalidade do *Big Data* proporciona uma nova visão sobre os dados não tradicionais para a análise preditiva, em que bases de dados quando associadas tem um grande valor, mesmo que seus dados individualmente forneçam pouca informação.

O *Big Data* possuí um grande leque de atuação. Sua tecnologia pode ser utilizada para os mais diversos ramos para basear decisões seguras e eficientes. Por exemplo, pode ser utilizada para segurança de eventos, segurança na infraestrutura de TI, melhorar ações de *marketing*, prever ações de mercado, etc. Existem casos do seu uso no desenvolvimento de aplicativos, no mercado esportivo e até mesmo para auxiliar a polícia criminalmente. É usado para decidir contratações, em decisões de crédito e pesquisas acadêmicas. Há uma infinidade de possibilidades para os mais diversos mercados e interesses.

Isso é reforçado por Hoerl et al. (2014) que diz que o *Big Data* chegou para ficar e está revolucionando as mais diversas áreas, como negócios, governos, indústrias, educação, medicina, ciências e, potencialmente, todos os aspectos da sociedade. Isso exige ainda que profissionais de diferentes áreas, como estatísticos, cientistas e engenheiros, além de outros, trabalhem suas habilidades em conjunto para conseguir lidar com o *Big Data*.

Para um melhor entendimento de *Big Data*, é necessário realizar uma análise de suas principais dimensões. Nos trabalhos sobre *Big Data*, no geral, são encontrados 5 critérios:

Volume, Variedade, Valor, Velocidade e Veracidade. A Figura 7 abaixo representa essas dimensões.

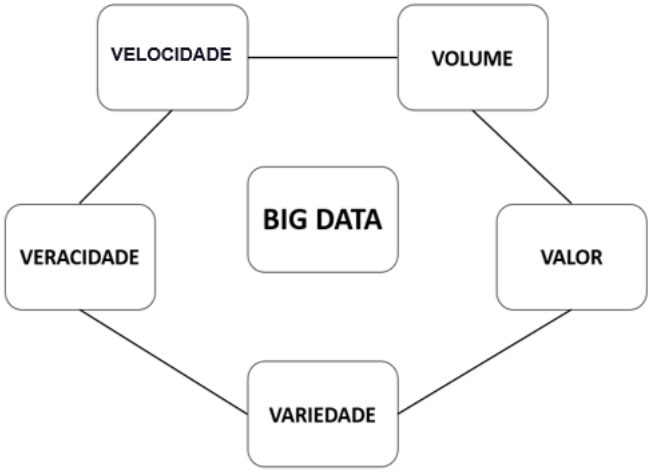

Figura 7 – Dimensões do *Big Data*

Fonte: própria.

- Volume

Este critério diz respeito à grandiosidade e extensão dos dados, e se destaca pela esmagadora quantidade de informações nas bases de dados do *Big Data* em comparação ao outro tipo. Esse volume é tão alto que no universo do *Big Data* já são utilizados termos como Petabytes e Exabytes. Um Petabyte equivale a 10^6 Gigabytes. Já um Exabyte é igual a 10^3 Petabytes. Com isso pode-se ter uma ideia do tamanho da grandeza utilizada atualmente para a obtenção de dados.

Boyd e Crawford (2011) afirmam que muito do entusiasmo que o *Big Data* gera está relacionado ao fato de que ele oferece a percepção de um acesso fácil para uma quantidade imensa de dados.

Martin (2015) acrescenta que para o desenvolvimento do estudo com o *Big Data* é necessário que se tenham conjuntos de dados grandes, complexos e vindos de diversas fontes. O aumento da quantidade de dados auxilia esse processo de avanço. Por exemplo, uma máquina que aprende a partir de algoritmos se baseia nos dados para a aprendizagem, e quanto mais dados forem introduzidos, maior será o aprendizado da máquina, criando um ciclo virtuoso (LOHR, 2012).

Desafios quanto a esse aspecto do *Big Data* estão relacionados ao armazenamento e processamento dos dados. O volume de dados é tão grande que a maioria dos *softwares* comerciais de análise de dados não conseguem processá-los (HOERL et al., 2014). O *Big Data Analytics* fornece ferramentas para lidar com essa grande quantidade de dados e as dificuldades associadas a isso. Cook e Forzani (2018) ainda dizem que esses problemas, que geralmente surgem da ideia de que tantos dados não cabem em um armazenamento tradicional, devem ser armazenados, gerenciados, tratados e analizados através de *clusters* ou na nuvem.

- Variedade

Quanto ao critério Variedade, o *Big Data* apresenta seus dados de forma muito diversificada, em dados estruturados, não estruturados e semi estruturados enquanto os dados tradicionais costumam ser estruturados.

Gandomi e Haider (2015) dizem que é estimado que apenas uma pequena parte *da Big Data* seja formada de dados estruturados e já prontos para a análise e que, a grande maioria dos dados é no formato não estruturado e parcialmente arquivados. Estes dados não estruturados, são dados que geralmente não possuem a organização estrutural que as máquinas que realizam a análise de dados necessitam. Texto, imagens, vídeos e áudios são alguns exemplos desse tipo de dados.

Segundo Russom (2011), o *Big Data* possui tipos de dados ecléticos, algo nunca visto antes, uma mistura de dados estruturados, não estruturados e semi estruturados, além de dados multidimensionais que podem ser extraídos de alguma base de dados para adicionar contexto histórico aos dados, sendo o aspecto variedade tão grande quanto o volume no *Big Data*.

Tendo em vista que grande parte do *Big Data* surge de forma não estruturada, o que não é adequado para os bancos de dados tradicionais, faz-se necessário o uso de ferramentas que consigam adquirir conhecimentos a partir destes dados não estruturados, que estão ganhando cada vez mais território e avançando em suas técnicas. A grande variedade de dados não é nenhuma novidade, o aspecto inovador da *Big Data* se deve ao surgimento de novas tecnologias na análise e gerenciamento de dados (GANDOMI e HAIDER, 2015).

Devido à grande variedade e volume de dados é difícil coletar e integrar essas informações em escala. Além disso, os sistemas de *Big Data* precisam armazenar os dados e gerenciar suas bases de dados enquanto garantem uma boa performance de informações e

protegem a privacidade dos dados. As análises das bases de dados devem, ainda, ser feitas em tempo real (HU, 2014).

- Valor

Valor diz respeito à importância e relevância que se consegue extrair dentro de uma grande variedade de métodos existentes para a análise dos dados do *Big Data*. Ele pode ser obtido e gerado mesmo que, dentro de um volume enorme de dados, se tenha uma baixa proporção de informações valiosas, pois o valor pode ser obtido através de *insights* que não são percebidas diretamente nos dados. Isso é confirmado por Gandomi e Haider (2015), que dizem que os dados na sua forma original possuem um baixo valor em comparação a seu volume.

Akter et al. (2017) cita que apesar da grande adoção do *Big Data* atualmente, muitas empresas ainda ainda enfrentam problemas em conseguir informações de qualidade dos dados. Por isso, ele define *quality analytics Big Data*, como a excelência da plataforma de *Big Data* percebida pelos usuários, ou seja, pela capacidade da plataforma de produzir informações de qualidade para os negócios. Além da grande variedade de áreas em que o *Big Data* pode atuar os seus dados também vem de variadas fontes. O *Big Data* combina essas informações diferentes de formas inovadoras para criar conhecimento.

Hoerl et al. (2014) reforça essa ideia ao dizer que um bom entendimento de onde vieram os dados e de como foram coletados fornece uma melhor percepção da qualidade e valor destes dados, se tornando essencial para a análise dos dados entender o processo que gerou esses dados.

Gerar valor que seja útil para organizações e para a sociedade é um dos grandes desafios atuais do *Big Data*.

- Velocidade

O critério Velocidade refere-se à taxa de rapidez em que os dados são gerados e a grande velocidade de processamento e análise destes. Isso traz novos desafios à análise de dados, pois os padrões desejados e percebidos estão mudando constantemente, o que não ocorre com dados estáticos (ABBASI et al., 2016).

Gandomi e Haider (2015) destacam que as novas tecnologias possibilitam uma geração de dados tão grande e rápida que tem o potencial de ser usada para oferecer em tempo real opções personalizadas de produtos e serviços para diferentes clientes. Segundo Russom (2011)

a análise desses tipos de dados é desafiadora, pois precisa encontrar sentido nos dados e possivelmente realizar ações em tempo real.

- Veracidade

Veracidade diz respeito à confiabilidade das informações obtidas e das fontes usadas. Um dos grandes problemas enfrentados atualmente é a grande quantidade de *spam* existente nos dados obtidos na internet. Esses problemas exigem que se tenha um gerenciamento, tratamento e conhecimento de dados apropriado.

Outro ponto relevante é que o *Big Data* pode ser originado de fontes desconhecidas ou questionáveis, por isso, é esperado que quem utilize os dados convença as pessoas de que o *Big Data* utilizado foi gerado de forma válida e apropriada (AGARWAL e DHAR, 2014). Desta forma, a necessidade de lidar com dados incertos e imprecisos é mais um dos desafios enfrentados pelo *Big Data*.

De uma maneira geral, todos esses critérios são responsáveis pela diferenciação *do Big Data*, e dependem de fatores que não possuem limitações definidas ainda. O principal que deve ser entendido é que esses critérios são dependentes um do outro, e que devem ser analisados em conjunto para o melhor estudo do *Big Data*.

Martin (2015) cita que apesar dos benefícios, o *Big Data* também apresenta usos questionáveis, como a falta de privacidade dos dados e o uso potencialmente discriminatório dos dados. Ele alerta também que o *Big Data* muitas vezes é usado como moralmente neutro e tendo benefícios que superam qualquer custo. Além disso, o uso irresponsável do *Big Data* pode intensificar desigualdades econômicas, desestabilizar mercados globais e reafirmar tendências preconceituosas (DROSOU et al., 2017).

Outro ponto é a necessidade de garantir que a aplicação do *Big Data* realmente leve para um caminho de economia desejável e resultados que sejam relevantes para os *stakeholders*. Agarwal e Dhar (2014) dizem que, apesar de encorajarem o estudo de temas que já foram explorados, pois os testes e replicações de pesquisas já realizadas são essenciais para o desenvolvimento da ciência, os resultados e pesquisas que serão considerados mais relevantes são os que escolherem fazer questionamentos novos, interessantes e relevantes para a sociedade.

Debates sobre o *Big Data* questionam as limitações dos modelos estatísticos tradicionais, que falham em capturar relações detalhadas entre indivíduos e grupos enquanto eles mudam em determinados contextos. As técnicas vistas no *Big Data* permitem capturar e

analisar dados em variadas formas e advindos de redes que possuem relações complexas entre suas entidades (ARADAU e BLANKE, 2017).

É um erro presumir que uma quantidade gigantesca de dados mais ferramentas avançadas de *analytics* significam necessariamente sucesso. Em muitas situações encontram-se problemas nessa fórmula. Dados que não possuem relevância para a pesquisa, em grandes quantidades, podem dificultar o processo de encontrar soluções. Hoerl et al. (2014) alertam que o problema não são as análises realizadas, mas sim nos dados e na forma como as análises estão sendo aplicadas, pois muitas vezes, devido à promessa dos avanços nas ferramentas e dados, acaba-se negligenciando fundamentos da estatística.

A aplicação dos princípios da estatística, como a qualidade dos dados utilizados e o uso de estratégias sequenciais para a aplicação de dados volumosos, complexos e não estruturados são essenciais para que o *Big Data* se aproxime do que é esperado dele. Ao se associar conhecimento, bons dados e um bom sistema de *analytics*, os resultados tem tudo para serem extraordinários para a sociedade (HOERL et al., 2014).

3.2 Uso dos dados

Embora a realidade do uso de *Big Data* não esteja alinhada em todo mundo, as organizações têm investido no intuito de melhorar suas análises de dados. Muitas estão implementando formas específicas de análises mais avançadas, como análises preditivas e prescritivas. Essas técnicas existem há muito tempo, mas seu uso se destaca nos dias de hoje pela quantidade cada vez maior de empresas que as utilizam (RUSSOM, 2011).

A necessidade de estar em constante aprimoramento das organizações transcendeu do objetivo de conhecer seu ambiente para predizer e prescrever por meio de modelos. A análise de dados ajuda a descobrir o que mudou e como se deve reagir a essas mudanças.

3.2.1 Modelos Preditivos

Predição está relacionado a se ter algum tipo de orientação em relação ao futuro. A predição tem sido constantemente utilizada na história, através dos mais diversos meios, para

auxiliar as tomadas de decisões e intervenções de empresas, governos e pessoas (ARADAU e BLANKE, 2017).

Modelos preditivos são modelos que identificam padrões e relações em dados e, desta forma, oferecem uma previsão do assunto desejado, apresentando possibilidades futuras. Essa previsão oferece embasamento para tomadas de decisões de organizações de acordo com seus objetivos e estratégias. O resultado da análise preditiva são diferentes predições e suas respectivas probabilidades de ocorrência.

O objetivo dos modelos preditivos é, a partir da aplicação de modelos que utilizam amostras de dados, a realização de previsões para situações que não fazem parte desta amostra (SHMUELI et al., 2016).

Utilizando estatística, como modelos de regressão, por exemplo, mineração de dados, dados históricos, aprendizado de máquina, inteligência artificial ou outros tipos de algorítmos, a análise preditiva tenta prever possíveis acontecimentos em situações com características diferentes das já ocorridas. Uma observação importante é que para a obtenção deste modelo é necessária a realização de análise de dados históricos e atuais. Além disso, por fazer uso de ferramentas estatísticas, um ponto muito importante destes modelos são a confiabilidade, que deve ser alta.

Gandomi e Haider (2015) dizem que as técnicas de análise preditiva podem ser classificadas em dois subgrupos: primeiro, as que pretendem descobrir padrões históricos para os replicarem em situações futuras; no segundo, as que procuram registrar interdependências entre os dados, e a partir das explicações encontradas realizar as predições.

Segundo Junqué de Fortuny et al. (2013) “a modelagem preditiva é baseada em uma ou mais instâncias de dados para o qual se quer prever o valor de uma variável de destino”. Soltanpoor e Sellis (2016) definem que a análise preditiva identifica oportunidades e riscos através da distinção de padrões e que, para resultados mais acurados, quanto mais dados se tiver disponíveis melhor.

A aplicação deste modelo permite que sejam percebidos padrões que não seriam observados normalmente pelas empresas. Isso dá a elas a opção de atuar em situações com uma maior precisão aumentando a chance de sucesso. É possível prever o comportamento do mercado, dos clientes, da concorrência e com isso, aumentar a competitividade tomando ações cada vez mais seguras e específicas a cada situação. Segundo Shmueli et al. (2016) é

fundamental para a análise preditiva uma habilidade de prever informações mensuráveis sobre novos casos.

Existem dois tipos de modelos preditivos, os supervisionados e não supervisionados. No modelo supervisionado ocorre a entrada e a saída de dados simultaneamente, com o objetivo de fazer com que o modelo aprenda a encontrar as relações e identificar padrões entre os dados de entrada e saída, deste modo deve-se estabelecer a variável dependente. Já no não supervisionado ocorrem apenas a entrada de dados inicialmente, e a partir destes dados são procurados padrões, para depois gerar a saída de dados, ou seja, sem necessidade de estabelecer uma variável dependente.

3.2.2 Modelos Prescritivos

Saber o que aconteceu no passado e previsões para o futuro não é mais suficiente para se obter vantagens competitivas. É necessário transformar todos esses conhecimentos obtidos com a análise de dados em recomendações, decisões e ações que tenham valor para os interessados. Nesse contexto surge a análise prescritiva de dados, que se preocupa com a orientação que será dada para as organizações através de ações adaptáveis e automatizadas (SOLTANPOOR e SELLIS, 2016).

Modelos prescritivos são modelos que utilizam estatística associado com gestão para basear decisões tornando as estratégias das empresas mais eficientes. Ele visa não apenas identificar quais eventos podem ocorrer, mas as consequências dos eventos e os comportamentos que ele pode gerar. Busca, desta forma, encontrar os melhores resultados possíveis sugerindo ações que tragam vantagens e minimizem riscos.

A análise prescritiva geralmente possuí duas características significativas: oferece prescrições em termos de ações e disponibiliza mecanismos de feedback para buscar recomendações de melhoria e a ocorrência de eventos não esperados (SOLTANPOOR e SELLIS, 2016).

Apesar de tanto a análise prescritiva quanto a análise preditiva fazerem uso de ferramentas estatísticas, a diferença entre elas é que a prescritiva fornece informações para a tomada de decisão, identificando consequências para possíveis ações tomadas, desta forma fazendo uma prescrição e não somente uma previsão futura.

Em um contexto de *Big Data*, modelos preditivos e prescritivos, é normal que novos desafios se manifestem. Dispositivos móveis e suas inúmeras aplicações tem transformado a sociedade e a forma como as pessoas se relacionam com dados. A esmagadora quantidade de dados vindos da web e aparelhos móveis possibilitam novas descobertas, insights e informações cheias de detalhes e conteúdos relevantes para qualquer tipo de negócio (CHEN et al., 2012). Muitas transformações foram percebidas nos mercados graças aos dados obtidos em *e-commerce*, redes sociais, plataformas de recomendação, etc. Diferentemente dos dados tradicionais, esses tipos de dados provindos da internet costumam ser menos estruturados e contém muitas informações de opiniões de clientes e informações comportamentais. Agora é possível observar e dimensionar o comportamento humano em uma escala global (AGARWAL e DHAR, 2014).

Shmueli (2017), explica que o cenário atual é formado pela união de duas tradições, uma do uso de grande volume dos dados por parte da Biologia, Engenharia e Medicina e outra do uso de dados comportamentais por meio da Psicologia, Administração, Marketing, Sistemas de Informação, Ciência Política e Educação. Deste modo pensar no *Big Data* e nos diferentes tipos de modelos que podem ser gerados, devem ganhar uma nova perspectiva: *Big Data* Comportamental (BDC).

3.3 *Big Data* Comportamental

O *Big Data* Comportamental "é uma grande base de dados multidimensional que captura ações e interações humanas e sociais em um novo nível de detalhe e é disponibilizada para empresas, governos e pesquisadores" (SHMUELI, 2017b). Ela oferece uma contribuição para outros métodos de coleta de dados que querem capturar intenções, emoções, sentimentos e pensamentos, tendo a intenção de transformar o entendimento dos indivíduos e da sociedade. O BDC tem um poder transformador que se baseia nessa riqueza de informações dos fenômenos sociais e humanos que antes eram imensuráveis. Segundo White e Breckenridge (2014) esses dados "têm o potencial de permitir que pesquisadores entendam o comportamental humano em um grau nunca antes visto".

Shmueli (2017b) exemplifica algumas características que demonstram o diferencial do *Big Data* Comportamental em relação a outros tipos:

> 1. A coleta do *Big Data* Comportamental pode mudar o comportamento dos objetos de estudo (sujeito).
>
> 2. O estudo e a análise do BDC pode prejudicar e por em risco o sujeito do estudo em formas intangíveis.
>
> 3. O estudo do BDC pode ser complicado pelo livre arbítrio.
>
> 4. BDC pode mudar com o tempo por si mesmo ou em resposta à pergunta ou à análise.

A partir das descrições obtidas acima, é possível perceber que o fator humano que existe no *Big Data* Comportamental tem o potencial de criar um grande impacto na análise, que pode ser alterada de diversas formas, desde a coleta dos dados ou até mesmo durante a análise deles. Em resumo, o sujeito humano é incluído no contexto junto com a questão de pesquisa, o pesquisador e os dados. O sujeito tem a viabilidade de alterar a pesquisa e os seus resultados, a qualidade dos dados, entre outras possibilidades, ações que antes eram limitadas ao pesquisador, único agente ativo da pesquisa até então.

Junqué de Fortuny et al. (2013) dizem que um dos principais aspectos dos bancos de dados comportamentais é que eles são dispersos, o que significa que em diferentes instâncias a maioria das características tem um valor de zero. Com isso, eles querem dizer que, ao se escolher um consumidor ao acaso, por exemplo, ele provavelmente não terá relação com a maioria dos comércios, ou não terá visitado a maioria dos lugares geográficos, nem a maioria de páginas que existem na internet, etc. Com o crescimento do banco de dados comportamentais a dispersão também irá aumentar. Isso se deve ao fato de que cada pessoa tem um limite de ações que podem ser realizadas, e quando se tem um grande número de possíveis ações que podem existir, a quantidade de ações que cada indivíduo pode realizar se torna cada vez menor em comparação com o número total.

Diferenciando o *Big Data* Comportamental e o *Big Data* tradicional, a distinção primordial é que o BDC captura as ações e interações das pessoas, que são repletas de sentimentos, intenções, decepções etc, enquanto o *Big Data* Tradicional utiliza dados inanimados, com coletas automáticas de sistemas provenientes da Engenharia. Assim, o BBC se difere do modelo tradicional, pois os sujeitos sabem que estão tendo dados coletados, eles podem até mesmo modificar seus comportamentos para evitar retaliações ou gerar resultados que, apesar de não serem a realidade, são os que eles gostariam de disponibilizar (SHMUELI, 2017b). O *Big Data* Tradicional abrange dados mais objetivos e que não sofrem esse tipo de interferência pelos sujeitos que estão fornecendo os dados, porém o atual cenário dos dados está carregado de interação humana.

Shmueli (2017a, 2017b) também diferencia o *Big Data* Comportamental do *Big Data* Fisiológico/Médico (BDF). Enquanto o foco do Fisiológico é em métricas físicas precisas, o BDC foca no impacto dos dados nos indivíduos, suas interações e ações. Ambos possuem o sujeito humano incluído na dinâmica além da figura do pesquisador como agente. No BDF os participantes de experimentos sabem do seu papel e tem interesse em fazer parte do estudo, já no BDC os participantes geralmente não sabem que fazem parte de um experimento. No BDC também prevalece o uso de modelos de variáveis latentes, como modelos de equações estruturais, que são ausentes nas BDF.

Outro aporte importante do BDC é sua interação com a pesquisa científica, obedecendo certo rigor metodológico característico da pesquisa científica.

Assim, Shmueli (2017b), apresenta o BDC custodiado pela pesquisa científica e formado por três eixos, a relação entre o pesquisador, a questão da pesquisa, os dados e o sujeito humano:

- Relação entre pesquisador e a questão da pesquisa

No BDC, ao contrário da pesquisa comportamental tradicional, pesquisadores dependem menos de suas experiências e literatura prévia. No que diz respeito ao uso de BDC para responder questões já existentes de pesquisa, "o pesquisador deve justificar porque as novas variáveis medidas podem operacionalizar constructos de interesse." (SHMUELI, 2017b)

Já no caso do uso de BDC que fazem novos questionamentos de pesquisa o desafio é na falta de literatura prévia. Este é um problema especialmente para estudos que tentam quebrar paradigmas e trazer novas questões, identificando novas teorias.

- Relação entre BDC e a questão da pesquisa

No BDC, graças a interação humana, muitas vezes ocorre uma distinção entre a unidade de observação (o que é medido) e a unidade de análise (ao que a questão de pesquisa se refere). Ocorre, também, um *trade-off* entre informação e privacidade, que gera essa distinção entre as unidades de observação e análise. Essas ações usualmente ocorrem com maior frequência nos estudos de BDC do que nos estudos comportamentais tradicionais devido às tecnologias e plataformas que geram o BDC e devido à forma que elas são usadas.

Outro ponto que ocorre com maior frequência no BDC do que nas pesquisas comportamentais tradicionais é que a amostragem dos dados geralmente fornece uma super-abrangência (*over-coverage*) ou sub-abrangência (*under-coverage*), ou seja, quando se

considera que a amostra representa a população na sua totalidade, mas isso não ocorre pois, na primeira situação se incluem elementos que não fazem parte da amostra e na segunda situação ocorre a omissão de elementos que compõem a amostra. A forma que os dados de BDC são adquiridos os tornam mais suscetíveis a esse tipo de erro.

- Relação entre pesquisador e BDC

Nesta relação, o BDC oferece dados maiores e mais ricos do que os tradicionais, e isso gera desafios na coleta, armazenagem, transmissão, computação e na execução da pesquisa. Além disso, por possuir dados mais complexos e sinais complexos, requer estudos e modelagens mais complexas além dos desafios que já existem na análise de dados comportamentais.

- Relação entre pesquisador e sujeito humano

A forma de coleta de dados do BDC exige, muitas vezes, uma aprovação por um comitê de ética. Esse tipo de aprovação já era exigido nas pesquisas comportamentais tradicionais, porém existem algumas diferenças na transição do contexto tradicional para o de BDC:

Uma delas é o fato de que muitos pesquisadores de BDC não são familiarizados com a ética dos estudos de sujeitos humanos.

Outra questão é que empresas que geram BDC não possuem uma exigência de ter essa aprovação pelo comitê de ética, mesmo quando realizam experimentos, pois suas técnicas não aparentam causar intervenções diretas na vida ou corpo dos sujeitos envolvidos.

A figura 8 a seguir resume as relações e questões envolvendo o BDC.

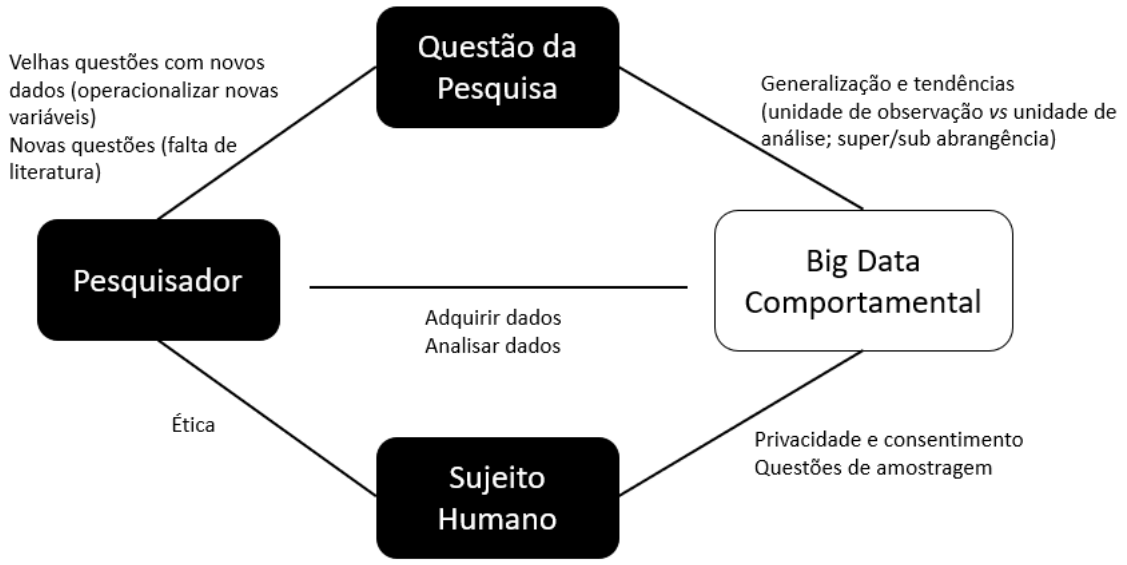

Figura 8 – Relações do *Big Data* Comportamental

Fonte: Shmueli (2017b)

Outro ponto é que no BDC existe uma grande distância entre o pesquisador e o sujeito da pesquisa, e isso gera uma dificuldade em prever riscos e danos e, desta forma, determinar o que deveria ser feito para garantir a segurança dos envolvidos. Além disso, os riscos no BDC costumam ser intangíveis, o que dificulta ainda mais esse processo.

Mais uma observação é que a aprovação pelo comitê de ética só é exigida em estudos considerados pesquisa. No contexto tradicional, a diferença entre pesquisa e prática é bem clara, porém, no contexto do BDC, essa diferença não é tão clara, pois mesmo que o pesquisador considere o estudo pesquisa, se for feita em colaboração com uma empresa já é considerada prática.

Com os avanços tecnológicos atuais, adquirir dados de BDC se tornou mais barato, mais rápido e mais automatizado (SHMUELI, 2017b).

A internet proporciona um maior acesso para dados comportamentais. Estes dados podem ser obtidos em diversas plataformas onlines, redes sociais, sites de relacionamento, e-commerce, sites de jogos, plataformas de busca, plataformas de pesquisas online e laboratórios virtuais, por exemplo. Esses dados geralmente contêm muitas e variadas informações sobre seus usuários, e, usualmente, esses dados de *Big Data* Comportamental são criados sem a intenção de gerar BDC, eles são gerados independentemente de serem solicitados ou não, e seu valor só é percebido depois (SHMUELI, 2017a).

Além da internet, outra forma de se coletar dados de BDC é através da Internet das Coisas, que se refere aos objetos físicos que possuem uma conexão com inteligência de computação, ou seja, uma competência de coletar e compartilhar dados. Esses objetos possuem a capacidade de coletar uma grande quantidade de dados comportamentais em ambientes e situações que antes não se tinha acesso para capturar dados, como alguns sistemas de monitoramento de ambientes de trabalho ou aparelhos inteligentes que aprendem os hábitos domésticos de seus usuários (SHMUELI, 2017a).

Dentro desse contexto de Internet das Coisas, Fawcett (2015) explica o conceito de *Quantified Self*, que se refere ao fenômeno das pessoas monitorarem vários aspectos da sua vida com a finalidade de conhecer, entender e otimizar seus hábitos, como o sono, quantidade de passos, batimentos, frequência de exercícios, etc. Dispositivos com essa finalidade são um dos maiores fenômenos tecnológicos dos últimos tempos, sendo responsáveis pela geração de uma grande quantidade de dados comportamentais. Ele destaca também que os usuários destes dispositivos esperam que eles gerem análises e *insights* de ações, ou seja, eles esperam que seja entregue uma análise prescritiva com esse tipo de dado comportamental.

Empresas também podem gerar dados de BDC, mas nem sempre possuem a capacidade de analizar esses dados. Em parceria com pesquisadores, elas podem usar estes dados para realizar pesquisas em temas que sejam do seu interesse, que forneçam informações que serão utilizadas para basear tomadas de decisões e gerar vantagem competitiva para a empresa. Podem, ainda, vender estas informações ou disponibilizá-las para a realização de pesquisas que não sejam diretamente ligadas a assuntos que gerem ganhos para a empresa, mas ainda seriam proveitosos para a sociedade.

É perceptível um progresso maior nas pesquisas de BDC quando realizadas por pesquisadores com um background misto, técnico, de engenharias, economia, ciência da computação e ciências comportamentais. O que diferencia essas áreas no uso do BDC é que as pesquisas são voltadas para alcançar objetivos de empresas, com perguntas de pesquisa sobre comportamentos humanos (SHMUELI, 2017a).

Assim como nas ciências sociais, a grande maioria das pesquisas no BDC são de natureza causal, inclusive para validar antigos questionamentos com os novos dados que são disponibilizados com o BDC (SHMUELI, 2017b). Contudo, ainda existem estudos que tem como objetivo prever comportamentos e a possibilidade de auxiliar na tomada de decisões, identificando novas questões que devem ser perguntadas e pesquisadas. De uma forma geral,

as pesquisas costumam focar em estudos causais enquanto as empresas tendem a preferir análises preditivas.

Estes estudos preditivos geralmente são realizados por pesquisadores com background em aprendizagem de máquina. Um exemplo de estudo que envolve modelos preditivos e BDC é o *target marketing*, que ajuda a entender quem é o público alvo de empresas, para disponibilizar ofertas customizadas ou descontos, entre outras possibilidades (SHMUELI, 2017a). Esses estudos são, muitas vezes, para basear decisões que devem ser tomadas rapidamente, no curto prazo, que exigem ações imediatas (SHMUELI, 2017b).

Outra observação sobre estudos preditivos com BDC realizado por Junqué de Fortuny et al. (2013) diz que o valor da predição cresce de acordo com a quantidade de dados comportamentais, portanto, empresas que possuem maiores bancos de dados comportamentais tendem a ter uma vantagem quando se trata de estudos com questionamentos preditivos.

3.3.1 Desafios no *Big Data* Comportamental

A partir da abordagem de Shmueli (2017a), que divide os problemas encontrados no BDC em categorias, também dessa forma será feito neste trabalho para um melhor entendimento. Os problemas são divididos em: técnicos – dificuldade de acesso aos dados, mudanças de ambiente, etc; de questões legais, éticas e morais – que envolvem privacidade dos dados, segurança dos sujeitos, conflitos entre as partes, etc; e de metodologia – questões sobre representatividade no uso dos dados, contaminação de dados, problemas estruturais, etc

- QUESTÕES TÉCNICAS

Um grande desafio do estudo da BDC é a falta de literatura prévia. Isso pode ser percebido especialmente na realização da revisão de literatura, que mostrou a quantidade pequena de textos que abordavam o assunto diretamente.

Nos últimos anos, o BDC tem sido disponibilizado de forma crescente, por entidades governamentais e empresas. Apesar disso e das facilidades conseguidas com os avanços tecnológicos e pela internet, os dados nem sempre são encontrados de maneira que permita

acessibilidade facilitada. Isso se deve ao interesse nas informações, ao formato em que as informações estão, regras de limite de compartilhamento, entre outros motivos. De uma forma geral, é possível obter dados de BDC, mas seu acesso ainda não é irrestrito e, apesar de crescente e de cada vez mais plataformas conseguirem obter dados, muitas ainda preferem não compartilhar seus dados de BDC.

Boyd e Crawford (2011) destacam que dificuldades na acessibilidade dos dados, como empresas que restringem o acesso, vendem o acesso aos dados por valores altos ou ainda oferecem banco de dados menores para pesquisas, podem produzir um desnível considerável no âmbito das pesquisas com BDC.

- QUESTÕES LEGAIS, ÉTICAS E MORAIS

Samarajiva e Lokanathan (2016) falam sobre como o surgimento do *Big Data* Comportamental possibilita o uso de algorítmos mais sofisticados que podem gerar comportamentos discriminatórios (não necessariamente ilegais) como mapeamento de riqueza, que permitem que prestadores de serviço encontrem mais facilmente seu público alvo e entendam melhor sua demanda, evitando práticas falhas, como, por exemplo, o uso apenas da localização para determinar mapeamento da riqueza.

Apesar de inicialmente este tipo de uso não levar a danos individuais, o avanço deste tipo de prática pode gerar. Por exemplo, pode-se ter discriminação através da recusa de organizações de atender pessoas em algumas regiões demográficas. Esse tipo de comportamento, apesar de poder ser mal visto pela sociedade, não é ilegal quando se trata de empresas, mas sim se realizado por entidades governamentais. Estas entidades podem, inclusive, se beneficiar dos mapeamentos ao verificar locais que necessitam de maiores investimentos.

Contudo, uma importante ressalva deve ser feita quanto a esse aspecto. Enquanto muitos benefícios podem ser obtidos com técnicas de predição de comportamentos, como produtos e serviços personalizados ou um conhecimento de onde se deve investir ou realizar ações sociais, etc, empresas que utilizam esse tipo de análise podem causar grandes danos sociais (SHMUELI, 2017b). Ao prever, por exemplo, a partir de um mapeamento de pobreza, pessoas que possuem um maior risco de inadimplência, isso pode gerar consequências sociais,

como recusa de empréstimos e de ofertas de trabalho, ajudando a criar práticas que acabam justificando e reforçando os comportamentos previstos, rotulando os indivíduos.

Acerca da segurança dos sujeitos envolvidos nas pesquisas de BDC, um dos pontos que deve ser observado é a privacidade dos sujeitos. Muitas vezes os dados, mesmo que estejam anônimos, podem identificar os indivíduos que o geraram, causando possíveis problemas para os envolvidos, tanto para os sujeitos quanto para as empresas. Segundo White e Breckenridge (2014) o cruzamento de dados provenientes de fontes variadas pode revelar informações dos indivíduos de maneira não intencional. Para Shmueli (2017a) essa integração do BDC entre plataformas variadas pode gerar riscos à privacidade das pessoas e, consequentemente, prejuízos para a sociedade no geral. Metcalf e Crawford (2016) também reforçam essa ideia ao afirmar que "dados disponibilizados publicamente podem ser usados de formas secundárias, incluindo a combinação com outros conjuntos de dados, podendo por em riscos indivíduos e comunidades".

Nos dias atuais a quantidade de dados públicos sobre os indivíduos é tão grande, que mesmo que pareçam inofensivos isolados, a junção deles se torna um risco à privacidade. De acordo com Berry (2011), sempre que se utiliza dados sobre sujeitos humanos existem dilemas quanto à privacidade, e é difícil quantificar os riscos de abuso de dados. Saramajiva e Lokanathan (2016) argumentam que dependendo de como os sujeitos percebem as situações como úteis ou incômodas é que surgem impasses quanto à privacidade.

A necessidade de se realizar análises em tempo real também acarreta problemas para a privacidade dos usuários, uma vez que tornar os dados anônimos ou pseudoanônimos é um procedimento a mais que toma tempo da análise (SARAMAJIVA e LOKANATHAN, 2016).

Pseudonimização é um modo usado para tentar proteger os usuários transformando um dado específico em um dado que não identifique o indivíduo, mas que, caso seja necessário, possa o tornar identificável. Isso pode ser feito ao se censurar os dados, utilizando iniciais ao invés de nomes completos, por exemplo, o que, dependendo do contexto do uso, também falha em garantir a privacidade do usuário.

Técnicas que permitem serviços de localização também podem ser utilizadas para rastrear movimentos de grupos ou indivíduos e alguns usos e propósitos desse tipo de informação podem trazer riscos. De uma maneira geral, quando se consegue relacionar as

pessoas no âmbito digital e no âmbito real, os danos relacionados à privacidade se tornam nocivos (SARAMAJIVA e LOKANATHAN, 2016).

Outro ponto levantado por Saramajiva e Lokanathan (2016) é sobre a privacidade de grupos. Algumas vezes, mesmo que o indivíduo esteja anônimo, é possível relacionar seus dados com outros e utilizá-los para obter informações sobre um grupo, como gênero, religião, localidades, universidades, empresas, etc. O impasse se dá quando a identificação do grupo relacionada a algum comportamento traz danos para o grupo, como julgamentos ou uma percepção da sociedade que desagrada as pessoas que fazem parte do grupo. Contudo, eles concluem que, apesar de ser possível a ocorrência de danos, nesses casos não é aconselhável que se aplique a mesma noção de privacidade do nível individual para o coletivo.

Como já foi dito anteriormente, Shmueli (2017a) fala sobre a necessidade de uma aprovação por um comitê de ética quando se coleta dados de BDC, pois se tratam de estudo que envolve sujeitos humanos, e é necessário assegurar que a pesquisa tem uma contribuição que justifique o risco que pode ter para as pessoas envolvidas. Práticas como dar o consentimento e proteger a privacidade dos indivíduos visam minimizar os riscos, tentar deixar a seleção dos sujeitos mais justa e prezar pela segurança, respeito e confidencialidade deles. Boyd e Crawford (2011) afirmam que, apesar de nem sempre ser possível prever os danos de uma pesquisa, o valor desses comitês de ética se dá ao fazer os responsáveis pensarem de forma crítica sobre a ética em seus estudos.

Um ponto que demonstra o porquê destes estudos não serem regulados como outros tipos é o fato de não se ter a percepção de que técnicas de *analytics* façam alguma intervenção na vida ou no corpo dos indivíduos. De uma maneira geral, os estudos nessas áreas têm um contato mais distante com os sujeitos humanos envolvidos, que são vistos como meios para se testar sistemas e não como o objeto de interesse em si (METCALF e CRAWFORD, 2016).

Contudo, como levantado por White e Breckenridge (2014), muitas vezes os pesquisadores não tem acesso e controle às atividades de coleta de dados, e o consentimento e autorização dos sujeitos usados na pesquisa pode não ter sido obtida da forma que geralmente é requisitada para as pesquisas. Mesmo assim, Boyd e Crawford (2011) reforçam que o processo de avaliação da ética no uso dos dados não pode ser ignorado simplesmente porque os dados estão disponíveis, e os pesquisadores devem continuar se questionando quanto à ética na obtenção, análise e utilização dos dados.

Quando se trata de dados públicos, que podem ser utilizados por pesquisadores, Boyd e Crawford (2011) ainda defendem que se deve levar em consideração a ética do uso, pois quando as pessoas publicam e disponibilizam seus dados, muitas vezes não imaginam que os dados podem ser utilizados para pesquisas, nem os riscos e benefícios que isso pode gerar. Tornar os dados públicos não significa necessariamente que foi autorizado seu uso para qualquer finalidade. Além disso, uma vez que os limites das pesquisas com sujeitos humanos ainda esta sendo muito discutido, é essencial que os estudos das ciências de dados se atentem para os potenciais riscos para os indivíduos se eles querem ter a confiança e apoio da sociedade em suas pesquisas (METCALF e CRAWFORD, 2016).

Ainda existe o problema em relação a possíveis conflitos entre o interesse das empresas e o interesse da pesquisa. Shmueli (2017b) diz que, geralmente, as empresas pretendem focar em questões do seu interesse e nos dados relativos a essas questões, que envolvem, por exemplo, predições de comportamentos de clientes, e é muito difícil existir um alinhamento entre os objetivos das empresas e acadêmicos. Os pesquisadores, devem, então, avaliar os impactos dos seus estudos, as suas responsabilidades em garantir a segurança e privacidade dos sujeitos e aliar o interesse da pesquisa com os interesses da empresa. O entendimento das implicações éticas do uso de dados, em termos de *trade-offs*, demonstra que pesquisadores já lidam com essas questões diariamente, embora isso não seja sempre óbvio para os cientistas de dados (BAROCAS, 2017).

Além disso, pesquisadores podem ficar inibidos em fazer alguns questionamentos se isso significar ir contra o interesse e pensamento de alguma empresa que possua o controle ao acesso de dados necessários para realizar as pesquisas. Esse efeito de suavizar as perguntas de pesquisa é perigoso para o futuro do BDC e para os progressos que podem ser obtidos através dos estudos (BOYD e CRAWFORD, 2011).

Outra situação que pode gerar conflitos entre a empresa e pesquisadores é quando o resultado da pesquisa é arriscados ou desfavorável para a empresa, o que pode fazer a empresa não autorizar a publicação. Uma forma de tentar evitar isso é tornar a empresa anônima no estudo (SHMUELI, 2017b).

- QUESTÕES DE METODOLOGIA

Abbasi (2016) destaca algumas questões que devem ser analisadas criticamente para que a análise de dados seja aproveitável. Uma delas é que se deve analisar o quão bem os dados representam uma população e quais interesses podem estar sendo excluídos ou super representados. Isso é um problema especialmente quando a análise de dados simplesmente assume que está representando a população completamente, o que muitas vezes não é o caso.

Ainda sobre representatividade, é necessário se considerar que muitas vezes a comunidade *online* não representa a totalidade da comunidade *offline*, especialmente em países em que a utilização da internet ainda é baixa pela população (SHMUELI, 2017a).

Junqué de Fortuny et al. (2013) também alertam para essa questão, por exemplo, quando se tem uma quantidade limitada de bancos de dados comportamentais e, devido a dispersão dos dados, pode ser que nem todos os comportamentos existentes sejam percebidos, mas apenas os contidos dentro daquelas instâncias limitadas, o que leva a uma não reprodução de todos os comportamentos existentes.

White e Breckenridge (2014) afirmam que o alcance de BDC sendo coletado ainda não é tão diverso quanto o tamanho de temas que são abordados nas ciências sociais, o que reforça esse entendimento que existem práticas e temas que não são representados na BDC. Segundo Boyd e Crawford (2011) “mesmo que um banco de dados tenha milhões de pedaços de dados, isso não significa que ele seja aleatório e representativo. É necessário entender as propriedades e limitações dos bancos de dados, por maiores que sejam.”

A diversidade, então, é um aspecto que muitas vezes é negligenciado quanto à responsabilidade no uso dos dados, que deve ser a garantia que tipos diferentes de sujeitos são representados em um processo de análise. Essa questão é importante não apenas pelo ponto de vista ético, mas também para garantir que as informações sejam precisas e atraentes (DROSOU et al., 2017). A amostragem dos dados é, portanto, um grande desafio do BDC, que se descuidado pode levar a conclusões erradas (SHMUELI, 2017a).

Outra questão que se destaca no caso do BDC é quanto ao efeito dos sujeitos humanos perceberem seu papel na pesquisa. Por exemplo, em experimentos com grupos focais, costuma-se utilizar técnicas, como placebo, para tentar evitar influências no estudo. Contudo, isso é mais complicado com BDC, pois os experimentos se dão em ambientes com alta conectividade, como é o caso da internet. Os sujeitos, muitas vezes conseguem se comunicar e percebem as manipulações ocorridas. Essa comunicação entre os sujeitos também faz com que sejam

compartilhados resultados ou informações do experimento, o que influencia na pesquisa. Isso exige criatividade dos pesquisadores para evitar esses efeitos (SHMUELI, 2017b).

Na estatística, é comum a utilização de análise de variância e modelos de regressão para testar relações de causalidade em dados experimentais, mas no BDC, quando se aplica esses métodos estatísticos minorias raras são filtradas do estudo e são ofuscadas pela maioria (SHMUELI, 2017a). Além disso, existem riscos de se observar padrões fictícios em quantidades massivas de dados, pois elas oferecem conexões em múltiplas direções. Portanto, é extremamente necessário que se leve em consideração as hipóteses analíticas, os *frameworks* metodológicos e os viés que podem existir no estudo dos dados (BOYD e CRAWFORD, 2011).

Uma vez que as técnicas das ciências de dados afetam demasiadamente as vidas das pessoas, as áreas encontram uma necessidade de um *framework* ético forte. A abordagem deve questionar como a subjetividade é construída nos conjuntos de dados usados nas pesquisas (METCALF e CRAWFORD, 2016).

Shmueli (2017a) comenta, ainda, que seria vantajoso existir uma metodologia que ligasse o objetivo de estudo do BDC com ferramentas adequadas para a análise destes dados. Um *framework* auxiliaria também quando ocorresse objetivos conflitantes, definindo princípios que devem ser seguidos para esse tipo de pesquisa. Relacionado a isso, Boyd e Crawford (2011) alertam que os resultados dos estudos com dados são mais efetivos quando se leva em consideração os métodos complexos que são exigidos para a realização da análise deste tipo de dados.

De uma maneira geral, os pesquisadores que utilizam dados comportamentais não costumam se basear em análises preditivas para basear seus modelos (SHMUELI, 2017b). Contudo, é perceptível a cada vez maior necessidade, não só das análises preditivas nestes dados, mas de análises prescritivas. E essa demanda surge tanto através dos indivíduos, que esperam que com o avanço tecnológico seus dispositivos se tornem mais inteligentes e façam recomendações a partir das análises que realizam, quanto de empresas e pesquisadores, que precisam dar indicações de ações a serem tomadas e para a melhoria da sociedade a partir dos dados comportamentais.

Os desafios encontrados na aplicação e uso do *Big Data* Comportamental são muitos, e a pesquisa sobre o assunto gera maiores debates e novos olhares sobre o tema. Solucionar esses desafios ainda é uma dificuldade comum aos trabalhos sobre o assunto. Possivelmente

com o uso cada vez maior desse tipo de dados e com o crescimento de trabalhos na área suas vantagens ficarão cada vez mais claras e óbvias para todos.

Porém, enquadrar o BBC dentro das premissas científicas requer o uso de um princípio metodológico de uma ou mais ferramentas que garantem a materialização do *Big Data* Comportamental.

4. METODOLOGIA

Neste trabalho foi utilizada metodologia exploratória por meio de pesquisa sistemática da literatura com abordagem qualitativa e quantitativa.

A metodologia exploratória geralmente é escolhida quando ainda há poucos estudos e conhecimentos sobre a temática abordada, o que é o caso deste. Espera-se, então, conhecer o assunto com maior profundidade e construir questões relevantes, auxiliando na construção de uma linha de estudo sobre a temática (RAUPP et al., 2006).

A pesquisa teórica teve como função embasar teoricamente o trabalho e todas as análises necessárias.

> A pesquisa bibliográfica é feita a partir do levantamento de referências teóricas já analisadas, e publicadas por meios escritos e eletrônicos, como livros, artigos científicos, páginas de web sites. Qualquer trabalho científico inicia-se com uma pesquisa bibliográfica, que permite ao pesquisador conhecer o que já se estudou sobre o assunto. (FONSECA, 2002)

A pesquisa bibliográfica possibilita novos olhares para determinado assunto, permitindo que se chegue a novas conclusões, sob novas perspectivas. Além disso, é a responsável pela fundamentação das ideias que serão apresentadas e desenvolvidas.

Com a abordagem quantitativa os resultados buscados podem ser quantificados, e existe a intenção de analisar as relações entre as variáveis.

> A pesquisa quantitativa se centra na objetividade. Influenciada pelo positivismo, considera que a realidade só pode ser compreendida com base na análise de dados brutos, recolhidos com o auxílio de instrumentos padronizados e neutros. A pesquisa quantitativa recorre à linguagem matemática para descrever as causas de um fenômeno, as relações entre variáveis etc. (FONSECA, 2002)

Já na abordagem qualitativa são realizadas análises mais profundas sobre o tema abordado, visando demonstrar aspectos que não foram observados pela análise quantitativa. A

pesquisa qualitativa foca no processo e no seu significado, e nela os dados analisados não são baseados em números. Esse método consiste na junção da observação e dos conceitos já formulados. Segundo Fonseca (2002), a junção entre os dois tipos de pesquisa, qualitativa e quantitativa permite que se tenha um acesso maior a informações, o que é vantajoso para a pesquisa.

4.1 Local de Estudo

Este estudo ao ser bibliográfico, foi realizado na plataforma *Web of Science*, sendo, portanto, uma representação dos registros mais relevantes do tema a nível mundial.

4.2 Objeto de Estudo

O objeto de estudo desta pesquisa foi o *Big Data* Comportamental.

4.3 Instrumento de Coleta de Dados

Para a coleta dos dados foi utilizada, principalmente, a base de dados *Web of Science*. *Web of Sciences* é uma base de dados de origem americana que abrange as mais diversas áreas de conhecimento científicos. É uma das bases de dados com maior alcance temporal, abrangendo publicações que chegam a 1900.

Além disso, permite que se realize uma avaliação da importância e influência de publicações, além de fornecer ferramentas de busca rápida, busca avançada, e busca por citação (FALAGAS et al., 2008).

Também fornece ferramentas que permitem analisar diversas informações sobre as publicações, como lugares onde as publicações foram realizadas, autores com maiores citações, idiomas em que as publicações foram realizadas, entre outras.

4.4 Tratamento de Dados

Para a realização do tratamento dos dados foi utilizada a ferramenta *VOSviewer*. O *VOSviewer* é um *software* gratuito, disponibilizado para a criação e visualização de mapas bibliográficos.

Segundo Van Eck et al. (2009), o *VOSviewer* permite que se mostre o mapa bibliográfico de diferentes formas, cada uma reforçando um aspecto que se prefira no mapa. Isto permite uma análise com maiores opções. No geral ele é indicado para usos com grandes quantidades de dados, mas seu uso é proveitoso no caso de poucos dados também.

5. ANÁLISES E RESULTADOS

Considerando-se um dos problemas apresentados sobre o *Big Data* Comportamental, no âmbito da metodologia, percebe-se a necessidade de uma metodologia e uma ferramenta que guie o pesquisador e o seu objetivo a uma forma de análise dos dados. A partir disto, buscou-se encontrar um método e uma ferramenta que estivesse de acordo com o significado de *Big Data* Comportamental e que pudesse potencializar seus resultados.

Dresch et al. (2015) explicam que a engenharia não se limita apenas a fenômenos e sistemas existentes, ela também se ocupa do projeto e estudo de fenômenos e objetos que ainda não existem, mas que serão criados pelo engenheiro para o bem comum, deste modo as metodologias tradicionais não atendem as necessidades das mudanças da engenharia, assim, como as novas demandas contemporâneas.

Os métodos tradicionais se limitam a explorar, explicar e, em algumas situações, predizer elementos e fenômenos existentes (Dresch et al., 2015). Nem sempre explicar uma situação e justificar sua ocorrência é o suficiente para resolver um problema e gerar conhecimento. A missão da engenharia é elaborar soluções que gerem contribuições e melhorias para a sociedade. No contexto da engenharia, muitas vezes é preciso lidar com situações que ainda estão sendo desenhadas, não existem.

Shmueli (2017 b) explica que o BDC é um dado diferenciado porque ele pode sofrer influência do comportamento humano. Além disso, é importante o entendimento de que o BDC

não pode ser tratado apenas como um dado comportamental e nem como um dado inanimado, mas sim como um constructo do comportamento humano em relação aos dados, e isso se aproxima do que é chamado de artefato.

Artefato, segundo Dresch et al. (2015) é “algo que foi construído pelo homem a partir de algum propósito. Uma interface entre o ambiente interno e o ambiente externo de determinado sistema”. Nesta definição, o ambiente interno caracteriza a forma como o artefato está organizado para atender aos objetivos exigidos e o ambiente externo define as condições de funcionamento do artefato.

Assim, com o avanço do *Design Science* (DS) para outras áreas, consolidando-se como *Design Science Research* (DSR), há as características da metodologia com o uso de artefatos. Além disso, existe na conjuntura atual, uma necessidade por estudos prescritivos, que indiquem caminhos e sugestões para basear decisões, o que também corrobora para a escolha da DSR. Os dados do BDC podem ser considerados artefatos. O DSR trabalha com uma perspectiva de artefatos e validação das informações, tendo, então, um uso possível no cenário do BDC.

Nos últimos anos há um aumento na utilização de “quase-experimentos”, ou seja, experimentos em que não se tem controle sobre todos os fatores, e isso torna necessário “o uso de técnicas mais robustas, automatizadas, que gerem *insights* e de fácil entendimento para os usuários” (SHMUELI, 2017b). Experimentos que utilizam o BDC são exemplos de situações em que não se tem o controle sobre tudo. E uma ferramenta que apresenta as características citadas é o PLS-SEM. Roldan e Cepeda (2018), explicam que entre as condições práticas de uso do PLS-SEM estão:

- Desenhos de pesquisa não experimentais (por exemplo, questionários, dados secundários, desenhos de pesquisa quase experimentais, entre outros) e se modelam grande número de variáveis latentes e manifestas

Cook e Forzani (2018) dizem que o PLS é um método interessante em conjunto com o *Big Data*, pois nele muitos preditores podem contribuir com informações relevantes.

Deste modo o uso do *Big Data* Comportamental na pesquisa em engenharias é possível, alinhado com o método *Design Science Research*, sendo operacionalizada por meio das Equações Estruturais via Variância (PLS-SEM).

5.1 *Design Science Research*

Segundo Dresch et al. (2015), o conceito do *Design Science* (ou ciência do projeto) nasceu das observações do prêmio nobel Herbert Simon, do que ele interpreta como Ciências do Artificial, livro lançado em 1969. Nesta obra o autor separa o que é natural do que é artificial em termos de objeto de pesquisa.

Neste novo contexto, o artificial é tudo aquilo idealizado/projetado pelo homem, como as máquinas, as organizações, a economia, e deste modo, as pesquisas que se ocupam do estudo de como criar e projetar novos artefatos, ou ainda, apoiar a resolução de problemas reais não se sustentam no paradigma das ciências naturais (DRESCH et al., 2015).

Aken (2004) descreve o conceito de ciências do *design*, que englobam engenharias, ciências médicas e psicoterapia moderna, e que tem como missão em suas pesquisas "desenvolver conhecimento válido e confiável para serem usados para conceber soluções para problemas". Deve-se enfatizar aqui que o objetivo não é a ação em si, mas o conhecimento gerado que poderá ser usado para fazer um *design*, um desenho, das soluções, e a partir daí sim que será realizada uma ação.

O *Design Science* é, então, um paradigma científico que surge para auxiliar pesquisas que tenham como objetivo a prescrição e, por consequência, a geração de conhecimento sobre como projetar (CAUCHICK, 2019). É uma alternativa de abordagem, que promove investigações através de artefatos que irão contribuir na criação de novos sistemas ou na melhora dos existentes, visando melhores resultados na resolução de problemas reais, se diferenciando das ciências naturais que apenas visam o estudo de fenômenos e problemas conhecidos, abandonando o conceito do "novo".

Assim, a ciência do artificial enxerga as ações que envolvem o homem como artefatos, uma conexão entre o ambiente interno (organização do artefato) e externo (condições de funcionamento), projetados para atender um determinado propósito, que devem ser validados e entregar uma solução satisfatória. Deste modo pensar nos modelos de *Big Data* Comportamental, é pensar na organização destes modelos e seu uso, com etapas claras de validação, o que ocorre quando o modelo tem sua fase de treino e teste, comum no *Data Science*.

Assim, no BDC, existe a necessidade de construção de um modelo que deve ser validado e desenvolvido etapa a etapa, camada por camada, similar ao contexto do *Design*

Science, que classifica e organiza os artefatos como constructos, modelos, métodos, instanciações e *design propositions* (CAUCHICK, 2019):

- Constructo: Conceitos utilizados para descrever problemas ou especificar respectivas soluções.
- Modelo: Conjunto de elementos e relações que representam a estrutura geral da realidade.
- Método: Conjunto de passos lógicos necessários para a efetivação de determinada atividade.
- Instanciação: Execução do(s) artefato(s) em seu ambiente real, evidenciando a viabilidade e eficácia dos artefatos.
- *Design Proposition:* Regras tecnológicas ou regras de projeto, consideradas contribuições teóricas da *Design Science.*

Deste modo, o *Design Science* se ajusta como paradigma suficiente para o *Big Data* Comportamental, garantindo que o aumento de pesquisas de caráter prescritivo nas engenharias contribua para o desenvolvimento do conhecimento científico. As pesquisas com abordagem da *Design Science* são orientadas à solução de problemas, sendo de natureza prática. Com o apoio da criação de artefatos, essas soluções visam beneficiar cada vez mais a sociedade, muito similar aos objetivos e preocupações do *Big Data* Comportamental.

Um ponto importante quanto ao *Design Science* é que a solução que ele desenvolve não precisa ser uma solução ótima, e sim uma solução satisfatória. Em outras palavras, significa dizer que o principal é resolver o problema, não sendo necessário que isso seja feito da melhor maneira existente, que pode ter impedimentos para a sua aplicação no mundo real. Segundo Dresch et al. (2015), o resultado é considerado satisfatório para a *Design Science* quando existe um consenso de que há um avanço da solução conseguida juntamente com o novo artefato em comparação a soluções anteriores, quando já existirem.

A partir da compreensão do conceito da *Design Science*, que é um paradigma, é possível o entendimento da *Design Science Research*, que é uma metodologia de pesquisa que aplica o *Design Science*. É, portanto, uma estrutura indicada quando se deseja prescrever soluções ou desenvolver e/ou avaliar artefatos (AKEN, 2004).

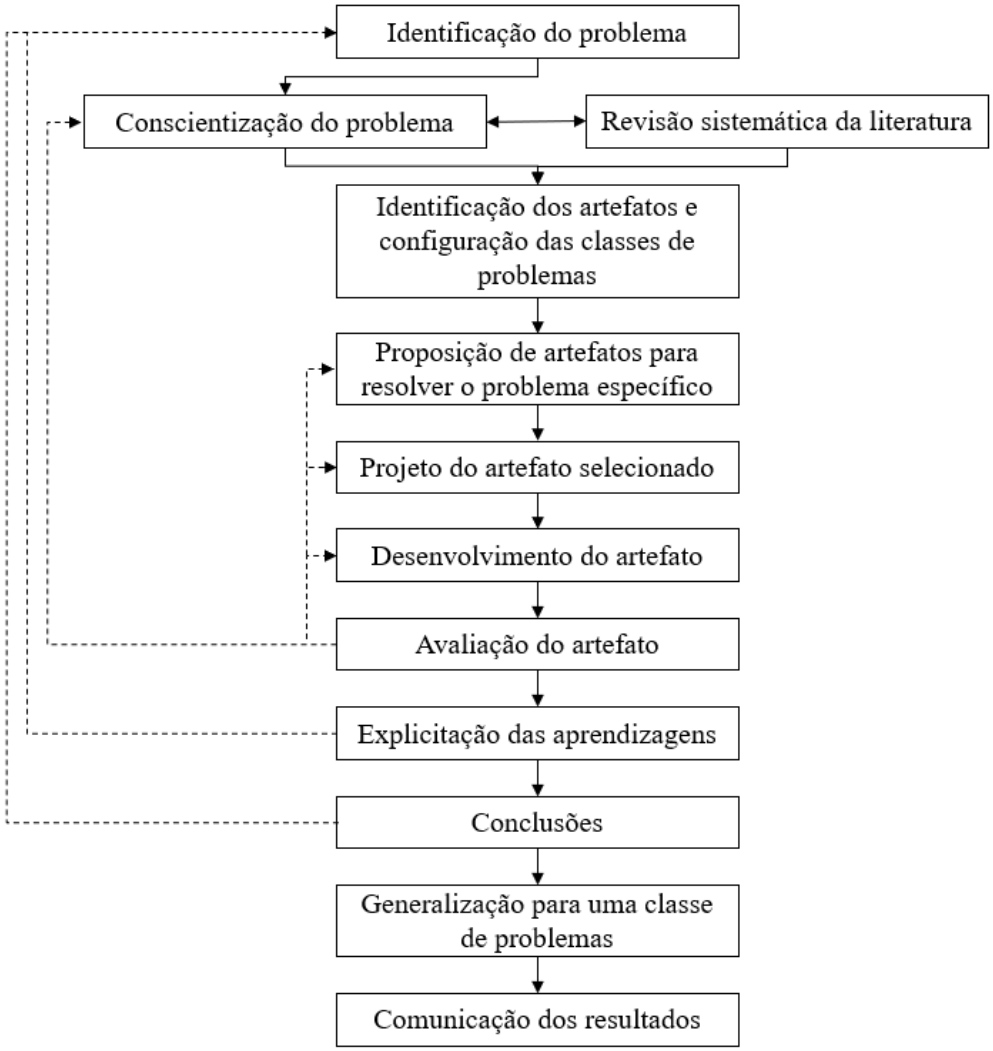

Figura 9 – Metodologia DSR

Fonte: Cauchick (2019)

Através da Figura 9, é possível verificar os passos necessários para a realização do *Design Science Research.*

A metodologia começa com a identificação do problema, que consiste em identificar, estruturar e compreender a situação para a qual se pretende apresentar uma solução. O entendimento do problema é essencial para que se possa criar uma solução útil e adequada. Em seguida deve-se realizar uma revisão sistemática da literatura, com o objetivo de fundamentar toda a pesquisa que será realizada, ou seja, embasar todas as ideias que serão apresentadas e utilizadas no trabalho.

O próximo passo é identificar possíveis opções de artefatos para serem desenvolvidos no intuito de solucionar o problema detectado anteriormente. Esta etapa costuma ser criativa e

subjetiva. Depois de se analisar as opções, deve-se, então, propor um artefato que se adeque ao proposto. Com o artefato definido, deve-se criar o projeto para desenvolver o artefato. Na *Design Science Research* o pesquisador também realiza o papel de projetista do artefato.

A próxima fase consiste na execução do projeto que irá resultar no desenvolvimento ou construção do artefato escolhido em seu estado funcional. O pesquisador constrói o ambiente interno do artefato, através de algoritmos, modelos gráficos, maquetes e outras ferramentas disponíveis.

Com o intuito de validar o artefato, o próximo passo é a avaliação do artefato. Depois de se definir quais requisitos devem ser avaliados e qual é a performance desejada do artefato, deve-se avaliar se o artefato cumpre o que se espera dele.

A explicitação das aprendizagens deve ser executada depois da avaliação, pois através dela é possível perceber as melhorias necessárias ao que foi realizado, evitando problemas no uso do artefato futuramente.

Para começar o processo de conclusão, relata-se tudo que foi ocorrido durante o trabalho. Depois disso, deve-se fazer, então, a generalização para uma classe de problemas. Vale citar aqui o conceito de "regra tecnológica" citado por Aken (2004), que é definido como "um conhecimento geral, que conecta um artefato com um resultado desejado em algum campo de aplicação". Ele diz ainda que uma regra tecnológica é um produto básico da *Design Science*, o que nada mais é do que a generalização para uma classe de problemas. Em outras palavras, esta etapa diz respeito ao fato de que a solução encontrada não é necessariamente apenas uma solução específica para um problema, mas um conhecimento que pode ser aplicado para uma classe de problemas semelhante ao enfrentado.

Por fim, comunica-se os resultados obtidos. Vale ressaltar que o trabalho precisa ser acessível e deve conter a descrição do que foi feito e de todos os resultados encontrados, incluindo problemas e limitações enfrentadas, pois essas informações irão auxiliar em novas pesquisas e em trazer novos conhecimentos.

Além da criação do artefato como resultado da *Design Science Research* para solucionar problemas, existe também a geração de conhecimento prescritivo, que cumpre o mesmo objetivo do artefato alinhado à metodologia com o *Big Data* Comportamental. Porém, o uso do *Design Science Research* abre um novo desafio, o de encontrar um instrumento que

execute a tarefa e que possa operacionalizar as prerrogativas do *Big Data* Comportamental em conjunto com as premissas do *Design Science.*

5.2 PLS-SEM

O incremento da tecnologia e a necessidade de melhorar o entendimento dos dados para a tomada de decisões tem ocasionado uma busca por métodos que possam projetar e entregar resultados aplicáveis. Neste contexto surgem as Equações Estruturais via variância (*Partial Least Square Structural Equation Modeling*) (PLS-SEM), uma união de duas importantes áreas, a econometria, por meio dos modelos preditivos e a psicometria, por meio dos modelos comportamentais.

Deste modo, a Modelagem de Equações Estruturais (SEM), surge como uma possibilidade de materialização dos princípios necessários para os modelos de BDC explicados por Shmueli et al. (2017).

A Modelagem de Equações Estruturais (SEM) é uma técnica que permite que se analise múltiplas relações entre fatores não observáveis e difíceis de mensurar, através de variáveis que funcionam como indicadores. Para avaliar as relações entre estas variáveis, são utilizadas equações, através de um conjunto de ferramentas estatísticas.

As variáveis representam constructos, que são os fatores não observáveis que se deseja analisar. Vale destacar que a definição do constructo é uma fase muito importante, pois erros na sua validação podem gerar danos em toda a pesquisa. O SEM, então, possibilita modelar e estimar parâmetros de relacionamentos entre os constructos.

Henseler (2017) define constructo como “construções que são justificadas teoricamente”, ou seja, um termo que descreve um acontecimento que se tenha um interesse teórico. Geralmente os constructos são formados por medições a partir de observação ou indicadores que representam os conceitos.

A SEM realizada através do método de *partial least squares* (PLS) se torna uma técnica que consegue estimar relações de causa e efeito entre constructos (HAIR et al., 2012). O PLS consegue representar conceitos não observáveis em modelos complexos. Segundo Shmueli et al. (2016) o PLS possui a habilidade de “produzir estimativas de parâmetros de

modelos complexos sem muitas das restrições distribucionais e outras questões dos métodos tradicionais".

O PLS-SEM é especialmente mais atraente quando o objetivo da pesquisa foca na predição e explicação da variância dos principais constructos alvos pelos diferentes constructos explicativos (HAIR et al., 2012). "A regressão através do PLS foi um dos primeiros métodos para predição em definições de alta dimensão nas quais o tamanho da amostra pode não ser grande em comparação ao número de preditores" (COOK e FORZANI, 2018).

Ramirez et al. (2014) descreve que o PLS deve ser usado em três fases: a primeira deve descrever o modelo estrutural, montando o gráfico do modelo e explicitando as relações causais entre as variáveis e os indicadores e constructos do modelo; a segunda fase é a validação do modelo, que verifica se os parâmetros utilizados estão dentro do intervalo aceitável, através do uso de cálculos estatísticos; e a terceira que é a valoração do modelo, que mostra o quanto as variáveis e o modelo conseguem explicar o problema inicial.

As características do PLS-SEM são aderentes ao contexto do BDC e *Design Science Research*, principalmente pelas descobertas realizadas nos últimos anos e explicadas no artigo de Henseler (2017). O principal achado foi observar os modelos de PLS-SEM desde a perspectiva de compostos.

Usualmente os constructos que trabalham comportamentos são tratados via fator comum, porém nos últimos anos se descobriu o modelo de compostos. Constructos de comportamento são, geralmente, variáveis latentes que podem ser compreendidas como entidades ontológicas, ou seja, da ciência do ser, como atributos de pessoas. Já os constructos de *design* são concebidos como produtos do pensamento. Enquanto os constructos comportamentais são ditos como fatores comuns, os constructos de *design* são chamados de compostos (HENSELER, 2017), o que é o mesmo que o artefato visto no DSR.

O SEM consegue estimar fatores comuns e compostos, o que faz com que ele seja aplicável tanto para constructos comportamentais quanto para os constructos de *design* (HENSELER, 2017).

Similar aos conceitos de *Design Science Research*, o PLS-SEM possui seus modelos ancorados em artefatos, que devem ser identificados na literatura, desenvolvidos, validados e aplicados, entregando resultados úteis.

A Figura 10 explica a interação entre o PLS-SEM e o *Design Science*.

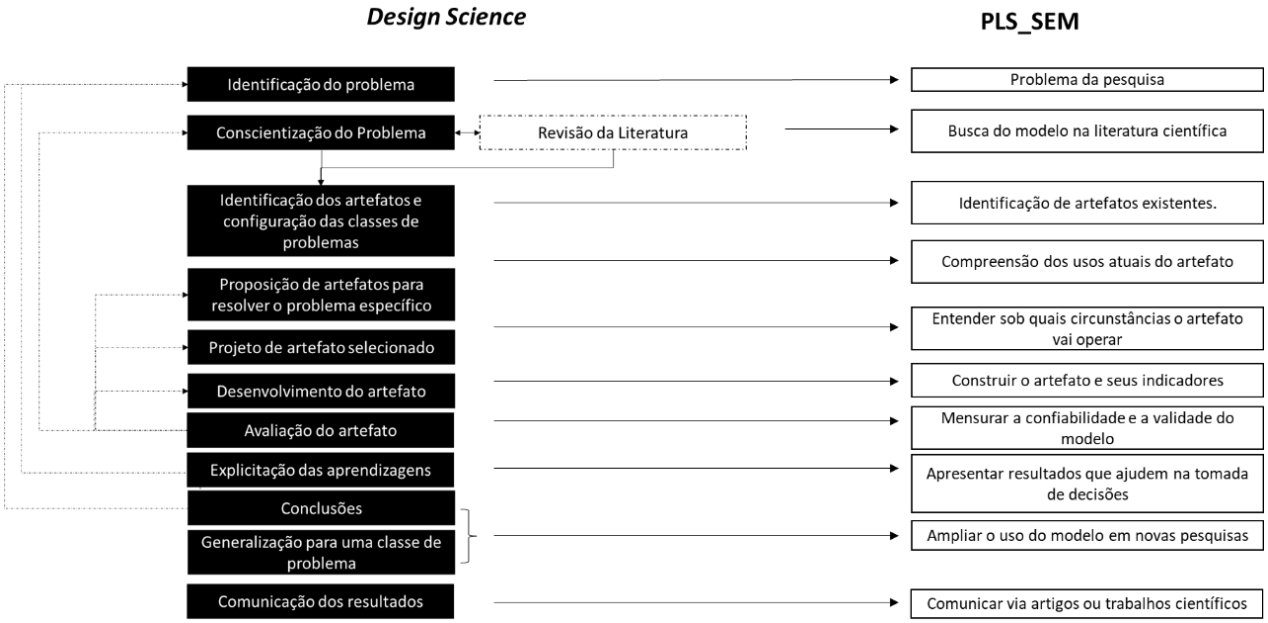

Figura 10 – Comparação entre DSR e PLS-SEM.
Fonte: Própria

Como é possível perceber, através da comparação vista na Figura 10, a parte inicial do uso do PLS-SEM e da DSR são iguais, começando pelo entendimento da questão de pesquisa e buscando o embasamento das ideias através da revisão de literatura, ou no caso do PLS, pela busca do modelo a ser utilizado na literatura. Essa semelhança não é uma surpresa, visto que é extremamente importante o entendimento do problema a ser estudado e o embasamento das ideias através da literatura, que garante confiabilidade ao trabalho através da fundamentação teórica.

A seguir, ambos executam a identificação de artefatos, um dos principais pontos que demonstram que o PLS aplica as ideias da metodologia DSR. A identificação de artefatos e a compreensão dos seus usos proporciona um entendimento da relação do artefato com a solução do problema da pesquisa. Essa fase também abrange a ideia da DSR de projetar o desenvolvimento do artefato.

A fase de construir o artefato e seus indicadores se assemelha com a fase de desenvolvimento de artefato da DSR, pois é a fase responsável pela execução do que foi projetado, resultando no desenvolvimento do artefato escolhido. Aqui hipóteses são propostas para se encontrar a relação entre os constructos e o problema inicial.

Através desses indicadores definidos, será possível realizar o próximo passo, que é a avaliação do artefato, que no PLS é a mensuração da validade e confiabilidade do artefato.

Nessa fase de avaliação, deve-se verificar se o artefato cumpre o proposto, descobrir suas limitações e se ele entrega o que é esperado.

Na fase final dos modelos, a metodologia DSR destaca a importância de se explicitar os resultados obtidos, o que deu certo e o que deu errado, pois tudo é importante para a geração de conhecimento a partir do que foi realizado. Nessa mesma linha, o PLS usa os resultados obtidos para auxiliar nas tomadas de decisões, com a apresentação do que foi encontrado.

A ideia de generalizar o problema para novas classes de problemas, ou seja, utilizar o que foi encontrado não apenas para solucionar o problema inicial, mas também tornar essa solução viável para outros problemas semelhantes, também existe em ambos os casos. O PLS destaca a importância de que o modelo criado seja utilizado para outros problemas similares, assim como o DSR propõe. Essa característica também é essencial para que o PLS seja considerado uma ferramenta que aplica a metodologia DSR.

Por fim, é esperado que todos os resultados encontrados sejam comunicados, ou seja, que sejam criados trabalhos científicos que compartilhem os conhecimentos com outras pessoas, o que possibilita que o uso das soluções e informações encontradas sejam disseminados para todos. Essa parte, assim como a inicial, não é surpreendente, pois é comum que isso seja esperado de trabalhos científicos que se propõem a resolver problemas.

Com o auxílio da Figura 10 e das explicações da metodologia DSR e da ferramenta PLS, além da comparação entre ambas, é possível perceber que elas se assemelham em seus aspectos principais, o que sugere que o PLS-SEM de fato é uma ferramenta que aplica as ideias propostas pela DSR.

Segundo Davenport (2014), o mundo usou mais de 2,8 zetabytes de dados (o que equivale a 2,8 trilhões de gigabytes) em 2012, porém apenas 0,5% destes dados são analisados de alguma maneira. As organizações e pessoas estão produzindo dados em tempo real, ocasionando em uma massa de dados que pode ser rica em informações para tomada de decisões.

Wang et al. (2015), explica que este avanço da quantidade disponível de dados não foi acompanhada pela transformação destes dados em informações úteis (...), e embora exista um aumento da tecnologia e das ferramentas disponíveis, este incremento não foi suficiente para garantir o uso destes dados para tomada de decisões. Este desafio aumenta quando os dados possuem formatos diversos e provém de conjuntura que envolve o ser humano.

O atual desenvolvimento das tecnologias provocou o uso de ferramentas e técnicas provenientes da estatística e das ciências de dados, porém esse avanço acabou por prejudicar uma série de passos necessários para garantir o rigor científico.

Todo uso indiscriminado da tecnologia gera efeitos. Laudon e Laudon (2004) abordam que ao introduzir uma nova ideia se cria um efeito concêntrico, como ondas, que irão abordar questões éticas, sociais e políticas, sucessivamente, e que serão tratadas nos níveis individuais, sociais e políticos. A Figura 11 mostra esse efeito.

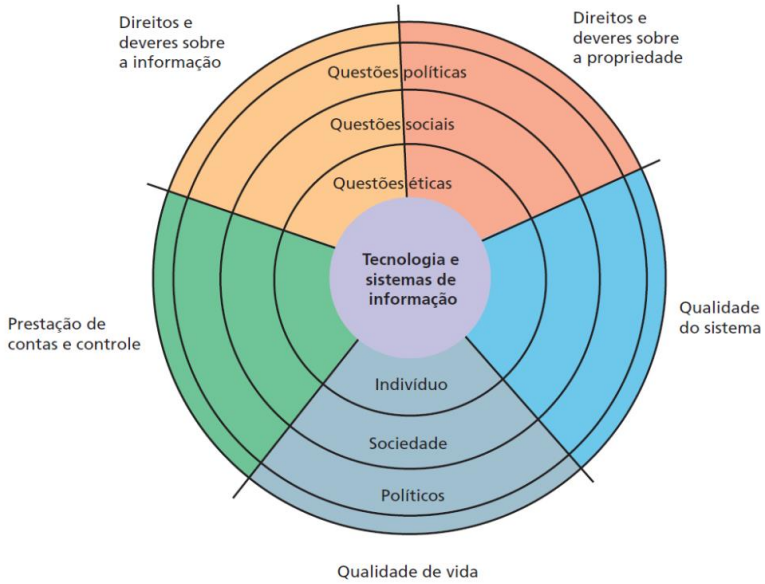

Figura 11 – Efeito de Ondas Concêntricas

Fonte: Laudon e Laudon (2004).

Assim, quando existe uma nova tecnologia e sistema de informação, o seu uso se dá inicialmente através do indivíduo e, consequentemente, alcança o nível da sociedade, o que é percebido no uso do BDC atualmente (Figura 11). Além dos diferentes níveis de uso, também existem as esferas de impacto relacionadas a cada uso, o que também já é visto no BDC, uma vez que já são identificados questionamentos quanto aos usos desse tipo de dado.

É importante ressaltar que quando o uso acontece primeiramente no nível individual é possível a existência de uma situação em que muitas vezes não se tem preocupação com questões éticas e sociais no uso das novas tecnologias e sistemas de informação. Contudo, com

a ampliação da esfera entra-se no nível de sociedade, o que gera discussões sobre os problemas observados na utilização da nova tecnologia.

Nesse contexto, é possível relacionar os desafios que são citados na Figura 11 com os encontrados para o BDC, como prestação de contas e controle, que pode ser exemplificada como a responsabilização dos danos causados pelos estudos comportamentais, e direitos e deveres sobre a propriedade, que pode ser relacionada à propriedade das informações, remetendo às questões já apresentadas, sobre privacidade de dados.

Com os problemas identificados nos níveis individuais e sociais, é necessário que se entre no nível político para que se tenha uma legislação que garanta o uso correto dessas novas tecnologias.

A Figura 11, então, pode ser aplicada, ao se entender que o BDC também funciona seguindo essa lógica, tendo questões que causam impactos em diferentes níveis, individuais, na sociedade e políticos, e portanto, ao se trabalhar questões no nível inicial, entende-se que esse efeito será aplicado às suas outras questões.

O uso de dados em grande volume também traz consequências e deve estar inserido no contexto de uso, principalmente na ciência. Assim como essa teoria diz, espera-se que, a partir da sugestão de uma metodologia e uma ferramenta, como foi apresentado, crie-se um efeito de ondas que faça a abordagem dos outros desafios que existem com o *Big Data* comportamental.

Assim, neste contexto, pode-se observar que o alinhamento entre *Big Data* Comportamental, *Design Science Research* e PLS-SEM são úteis e colaboram com o avanço de uso dos dados para a pesquisa científica, obedecendo os preceitos e rigor metodológico necessário.

6. CONSIDERAÇÕES FINAIS

Este trabalho tinha como problema explicar como o *Big Data* Comportamental poderia ser inserido no contexto da pesquisa científica na Engenharia de Produção. Para isso, primeiro era necessário delimitar o conceito de *Big Data* Comportamental e com isso identificar suas principais características e seus principais desafios.

A partir da análise realizada, percebeu-se uma necessidade de uma metodologia e uma ferramenta que combinasse com esse tipo de dados e suas características, pois, apesar de estudos com dados comportamentais já estarem sendo realizados, percebe-se uma falta de conexão metodológica entre os estudos. O objetivo do trabalho então, era apresentar uma construção metodológica para o uso do *Big Data* Comportamental na pesquisa científica na Engenharia de Produção.

A escolha do *Design Science Research* como metodologia se mostrou apropriada em especial pela sua característica de realizar estudos prescritivos e por ter uma abordagem de artefatos que engloba o BDC. Além disso, o PLS-SEM foi escolhido como ferramenta por fornecer uma combinação das ideias do DSR com o tipo de dado do BDC.

Desta forma, acredita-se que o uso do PLS-SEM a partir da metodologia DSR cumpre o papel de ser uma metodologia que oferece ao *Big Data* Comportamental o rigor metodológico exigido pelo estudo científico, ao mesmo tempo que proporciona para as organizações um método que potencializa o uso dos dados para análises prescritivas, que são cada vez mais uma exigência para a tomada de decisões, que devem ser cada vez mais rápidas e precisas.

Tendo em vista a Engenharia de Produção, o uso do *Big Data* Comportamental associado a essa metodologia e ferramenta é vantajoso, em especial pela perspectiva da Indústria 4.0 e seus desafios. O BDC também permite que se entenda melhor o comportamento do mercado e dos indivíduos e que, a partir disso, se conheça melhor os riscos e oportunidades envolvidos em cada situação. Além disso, com o uso do PLS-SEM associado à metodologia DSR, é possível verificar parâmetros e aspectos que impactam os comportamentos e ações, buscando pontos de melhoria e correção de erros. Para fazer as empresas crescerem e se tornarem competitivas globalmente, é necessário que se tenha tomadas de decisões rápidas e fundamentadas, e isso exige uma metodologia concisa e apropriada.

Uma das maiores limitações encontradas para o desenvolvimento deste trabalho foi a falta de publicações científicas sobre *Big Data* Comportamental que, apesar disso, deve ganhar cada vez mais espaço na literatura nos próximos anos devido à quantidade crescente de dados desse tipo que são gerados e do potencial do seu uso em pesquisas e no auxílio à tomada de decisão.

A partir do que foi demonstrado, próximos trabalhos podem fazer o uso da metodologia DSR com a ferramenta PLS-SEM para utilizar dados de BDC e verificar suas

vantagens para análises prescritivas. Além disso, acredita-se que a partir da abordagem do desafio metodológico encontrado no uso do BDC, se abram caminhos na abordagem das outras questões apresentadas. Com o maior uso do BDC nas pesquisas científicas será exigido que se apresentem soluções para os debates atuais existentes quanto à ética e moral no uso desses dados.

REFERÊNCIAS

ABBASI, Ahmed; SARKER, Suprateek; CHIANG, Roger HL. **Big data research in information systems: Toward an inclusive research agenda**. Journal of the Association for Information Systems, v. 17, n. 2, p. I, 2016.

AGARWAL, Deepak K.; CHEN, Bee-Chung. **Statistical methods for recommender systems.** Cambridge University Press, 2016.

AGARWAL, Ritu; DHAR, Vasant. **Big data, data science, and analytics: The opportunity and challenge for IS research**. 2014.

AKEN, Joan E. van. **Management research based on the paradigm of the design sciences: the quest for field-tested and grounded technological rules**. Journal of management studies, v. 41, n. 2, p. 219-246, 2004.

AKTER, Shahriar; FOSSO WAMBA, Samuel; DEWAN, Saifullah. **Why PLS-SEM is suitable for complex modelling? An empirical illustration in big data analytics quality**. Production Planning & Control, v. 28, n. 11-12, p. 1011-1021, 2017.

ARADAU, Claudia; BLANKE, Tobias. **Politics of prediction: Security and the time/space of governmentality in the age of big data**. European Journal of Social Theory, v. 20, n. 3, p. 373-391, 2017.

BAROCAS, Solon et al. **Social and technical trade-offs in data science**. 2017.

BERRY, David M. **The computational turn: Thinking about the digital humanities**. Culture machine, v. 12, 2011.

BOYD, Danah; CRAWFORD, Kate. **Six provocations for big data.** A decade in internet time: Symposium on the dynamics of the internet and society. Oxford, UK: Oxford Internet Institute, 2011.

CAUCHICK, Paulo et al. **Metodologia científica para engenharia**. Elsevier Brasil, 2019.

CHAUDHURI, Surajit; DAYAL, Umeshwar; NARASAYYA, Vivek. **An overview of business intelligence technology.** Communications of the ACM, v. 54, n. 8, p. 88-98, 2011

CHEN, Hsinchun; CHIANG, Roger HL; STOREY, Veda C. **Business intelligence and analytics: From big data to big impact**. MIS quarterly, v. 36, n. 4, 2012.

COOK, R. Dennis; FORZANI, Liliana. **Big data and partial least-squares prediction**. Canadian Journal of Statistics, v. 46, n. 1, p. 62-78, 2018.

DAVENPORT, Thomas H. **Big data no trabalho**. Elsevier Brasil, 2014.

DRESCH, Aline; LACERDA, Daniel Pacheco; JÚNIOR, José Antonio Valle Antunes. **Design science research: método de pesquisa para avanço da ciência e tecnologia.** Bookman Editora, 2015.

DROSOU, Marina et al. **Diversity in big data: A review**. Big data, v. 5, n. 2, p. 73-84, 2017.

FALAGAS, Matthew E. et al. **Comparison of PubMed, Scopus, web of science, and Google scholar: strengths and weaknesses**. The FASEB journal, v. 22, n. 2, p. 338-342, 2008.

FAWCETT, Tom. **Mining the quantified self: personal knowledge discovery as a challenge for data science.** Big Data, v. 3, n. 4, p. 249-266, 2015.

FONSECA, João José Saraiva. **Metodologia da Pesquisa Científica**. 2002.

GANDOMI, Amir; HAIDER, Murtaza. **Beyond the hype: Big data concepts, methods, and analytics.** International journal of information management, v. 35, n. 2, p. 137-144, 2015.

GOLDSCHMIDT, Ronaldo; BEZERRA, Eduardo; PASSOS, E. **Data Mining: Conceitos, técnicas, algoritmos, orientações e aplicações**. Rio de Janeiro-RJ: Elsevier, p. 56-60, 2015.

HAIR, Joseph F. et al. **The use of partial least squares structural equation modeling in strategic management research: a review of past practices and recommendations for future applications.** Long range planning, v. 45, n. 5-6, p. 320-340, 2012.

HENSELER, Jörg. **Bridging design and behavioral research with variance-based structural equation modeling.** Journal of advertising, v. 46, n. 1, p. 178-192, 2017.

HOERL, Roger W.; SNEE, Ronald D.; DE VEAUX, Richard D. **Applying statistical thinking to 'Big Data'problems.** Wiley Interdisciplinary Reviews: Computational Statistics, v. 6, n. 4, p. 222-232, 2014.

HU, Han et al. **Toward scalable systems for big data analytics: A technology tutorial.** IEEE access, v. 2, p. 652-687, 2014.

HURWITZ, Judith S. et al. **Big data for dummies**. John Wiley & Sons, 2013.

JUNQUÉ DE FORTUNY, Enric; MARTENS, David; PROVOST, Foster. **Predictive modeling with big data: is bigger really better?** Big Data, v. 1, n. 4, p. 215-226, 2013.

KRAMER, Adam DI; GUILLORY, Jamie E.; HANCOCK, Jeffrey T. **Experimental evidence of massive-scale emotional contagion through social networks.** Proceedings of the National Academy of Sciences, v. 111, n. 24, p. 8788-8790, 2014.

KUAZAQUI, Edmir. **Marketing in the New Times**. Journal of Marketing Management, v. 6, n. 2, p. 44-48, 2018.

LAUDON, Kenneth C.; LAUDON, Jane Price. **Sistemas de información gerencial: administración de la empresa digital**. Pearson Educación, 2004.

LIN, Mingfeng; LUCAS JR, Henry C.; SHMUELI, Galit. **Research commentary—too big to fail: large samples and the p-value problem.** Information Systems Research, v. 24, n. 4, p. 906-917, 2013.

LOHR, Steve. **The age of big data**. New York Times, v. 11, n. 2012, 2012.

MARIANO, Ari Melo; ROCHA, Maíra Santos. Revisão da Literatura: **Apresentação de uma Abordagem Integradora**. In: XXVI Congreso Internacional de la Academia Europea de Dirección y Economía de la Empresa (AEDEM), Reggio Calabria. 2017.

MARTIN, Kirsten E. **Ethical issues in the big data industry**. MIS Quarterly Executive, v. 14, p. 2, 2015.

METCALF, Jacob; CRAWFORD, Kate. **Where are human subjects in big data research? The emerging ethics divide.** Big Data & Society, v. 3, n. 1, p. 2053951716650211, 2016.

MINISTÉRIO DA INDÚSTRIA, COMÉRCIO E SERVIÇOS. **Agenda brasileira para a Indústria 4.0.** Disponível em <http://www.industria40.gov.br/>. Acesso em: 15 de fev. de 2019.

MOTA, Patrick; MARIANO, Ari Melo; MONTEIRO, Simone Borges Simão. **Taxonomy of the Industry 4.0: Theoretical and Practical Contributions to a New Context.** 2018.

NEGASH, Solomon. **Business intelligence**. Communications of the association for information systems, v. 13, n. 1, p. 15, 2004.

RAMÍREZ, Patricio E.; MARIANO, Ari Melo; SALAZAR, Evangelina A. **Propuesta Metodológica para aplicar modelos de ecuaciones estructurales con PLS: El caso del uso de las bases de datos científicas en estudiantes universitarios.** Revista ADMpg Gestão Estratégica, v. 7, n. 2, 2014.

RAUPP, Fabiano Maury; BEUREN, Ilse Maria. **Metodologia da Pesquisa Aplicável às Ciências. Como elaborar trabalhos monográficos em contabilidade: teoria e prática**. São Paulo: Atlas, 2006.

ROLDÁN, J L; CEPEDA, G. M1 – PLS-SEM, 5ª ed. – CFP. Universidad de Sevilla, 2018.

RUSSOM, Philip et al. **Big data analytics**. TDWI best practices report, fourth quarter, v. 19, n. 4, p. 1-34, 2011.

SANDERS, Na da R. **How to use big data to drive your supply chain.** California Management Review, v. 58, n. 3, p. 26-48, 2016.

SAMARAJIVA, Rohan; LOKANATHAN, Sriganesh. **Using behavioral big data for public purposes: Exploring frontier issues of an emerging policy arena**. 2016.

SCHWAB, Klaus. **A quarta revolução industrial**. Edipro, 2019.

SHMUELI, Galit. **Analyzing behavioral big data: methodological, practical, ethical, and moral issues**. Quality Engineering, v. 29, n. 1, p. 57-74, 2017a.

SHMUELI, Galit. **Research dilemmas with behavioral big data.** Big data, v. 5, n. 2, p. 98-119, 2017b.

SHMUELI, Galit et al. **The elephant in the room: Predictive performance of PLS models**. Journal of Business Research, v. 69, n. 10, p. 4552-4564, 2016.

SOLTANPOOR, Reza; SELLIS, Timos. **Prescriptive analytics for big data**. In: Australasian Database Conference. Springer, Cham, 2016. p. 245-256.

VAN ECK, Nees; WALTMAN, Ludo. **Software survey: VOSviewer, a computer program for bibliometric mapping**. Scientometrics, v. 84, n. 2, p. 523-538, 2009.

WANG, Yichuan; KUNG, LeeAnn; BYRD, Terry Anthony. **Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations**. Technological Forecasting and Social Change, v. 126, p. 3-13, 2018.

WHITE, Patricia; BRECKENRIDGE, R. Saylor. **Trade-offs, limitations, and promises of Big Data in social science research**. Review of Policy Research, v. 31, n. 4, p. 331-338, 2014.

XIAO, Bo; BENBASAT, Izak. **Designing warning messages for detecting biased online product recommendations: An empirical investigation**. Information Systems Research, v. 26, n. 4, p. 793-811, 2015.